Understanding the mechanism of racial bias in predictive risk models of child welfare

Huyen (Haley) Dinh

A dissertation submitted to

Auckland University of Technology

in partial fulfilment of the requirements for the degree of

Master in Business (MBus)

2021

School of Economics| Faculty of Business, Economics and Law

1. Abstract

Each year approximately 3.6 million children in the US are referred to Child Protective Services (CPS) – despite these high levels of surveillance, child maltreatment deaths have not fallen. Additionally, many children who are victims of abuse and neglect come to the attention of CPS when it is too late and where early intervention might have helped them. That is where Predictive Risk Modelling (PRM), a type of statistical algorithm that uses linked administrative data to predict the likelihood of adverse events happening in the future, comes into play. The PRM tool typically estimates a child's risk of abuse and neglect at the time of birth, then its predictions are employed to assist decision-making for connecting families to prevention services before incidents of abuse and neglect occur. However, there are growing concerns about racial disparity around the use of PRM in the child maltreatment context: whether it will reproduce, or even exacerbate, human bias. This study focuses on understanding one of the causes of machine bias, which is measurement error or target variable bias. In particular, the research investigates whether the use of a proxy variable, which is foster care placement in our context, can potentially lead to racial disparity in child maltreatment predictions.

Table of Contents

1. Abstract 2
2. Introduction 6
2.1 Predictive Risk Modelling (PRM) explanation 6
2.2. Outline of the dissertation 15
2.3. Data setup 15
3 What is Algorithmic bias? 16
3.1. Algorithmic bias and its causes 16
3.2. Obermeyer et al., (2019) 19
4. Research Questions 23
5. Definitions of algorithmic fairness 24
6. Fairness analysis for the current PRM tool 29
7. Mechanism of bias 46
8. Conclusion and Discussion 50
APPENDIX 52
References 61

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it does not include any material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher education.

Signature of candidate

Acknowledgement

Foremost, I would like to express my sincere gratitude to my supervisors Matthew Ryan and Rhema Vaithianathan, for their guidance, enthusiasm, and motivation. It was a privilege and honour for me to study and work under their supervision.

Especially, I am deeply grateful to Matthew Ryan, who always spends as much time as possible explaining everything that I did not understand. I want to give special thanks to him for being extremely patient with me. Matthew is the most devoted teacher by whom I am lucky to be taught. He pointed out to me my weakness when doing research. Nonetheless, he always encourages me when I am in doubt of my ability or my writing skill. Matthew has taught me how to write more logically and clearly and present my research work in a better structure. He also asks me many difficult questions and gives many insightful cues to help me understand better my research topic.

My sincere thanks also go to Rhema Vaithianathan for sharing her immense knowledge in the child welfare domain and econometrics with me. Even though Rhema has a hectic schedule, she is always there to support me and give me insightful comments.

Finally, I am deeply grateful to my family: my husband and my daughter, for always being my source of support. I could not have made it this far if it were not for their unconditional love, understanding and motivation.

2. Introduction

2.1 Predictive Risk Modelling (PRM) explanation

In recent years, machine learning algorithms - an application of statistical techniques that learn from historical data to make a future prediction - are increasingly applied to a wide range of practices in the private and public sectors. In the domain of child welfare, some agencies are supplementing manual child-at-risk detecting methods with automated decision algorithms that speed up the process of triaging the most vulnerable children and prioritizing them for protective services. It is crucial in child welfare systems to proactively identify families with complex needs to be able to help them thrive and to prevent child maltreatment from happening as early as possible. Although various preventive support programmes are given to families (e.g., home visiting), many of them have proven ineffective due to the lack of engagement by parents of children with the most complex needs(Vaithianathan et al., 2020). In other words, children and parents who are most in need of these services are not correctly triaged by the system, leading to missed opportunities to prevent adverse outcomes for children. Therefore, the system needs a means to better engage with the families before critical incidents occur. Machine learning algorithms, it is claimed, do a much better job of identifying children at risk of abuse and neglect than human judgment alone, due to their ability to process a large amount of information to produce highly accurate predictions (Drake et al., 2020). Moreover, these tools help shorten the assessment time and are able to triage the whole population of focused families, leading to greater efficiency and performance. For these reasons, machine learning algorithms are becoming increasingly popular in the child welfare domain.

Predictive risk modelling (PRM) is one form of machine learning algorithm that is currently used in the child welfare system. PRM tools are often trained on historical data using statistical methods such as LASSO (Least Absolute Shrinkage and Selection Operator), Random Forest and Support Vector Machines. These types of algorithms learn by constructing a set of rules that summarise the correlations between predictors and the target outcome on the training data set used to train the algorithm. Then it applies what it has learned to give predictions on the test set, which is unseen data not used in the training phase. In particular, PRM processes a large amount of administrative data collected and stored by government agencies and outputs the predicted probabilities of adverse events happening in the future. Like a threshold tool, it predicts that a child is at risk of maltreatment if their predicted risk probability is above a certain limit.

This study focuses on a preventive PRM tool used in the US' child welfare system. This preventive model utilizes the history of the child's parents and the child's birth record to predict the risk of out-of-home placement by age three. The data set used to train the model consists of 52,520 children born from 2012 to 2015 and having maternal residence in the jurisdiction. The predictors come from various domains such as Vital Birth Records, Children, Youth and Families, Jail and Court systems, Demographic data (excluding race), Homeless and Housing Support Services, and Juvenile Probation. The designers of the tool claimed that including race as an input variable was not useful in increasing the model's overall predictive power. They, therefore, excluded race and built the model from 459 predictors. However, the tool developers did not give evidence showing whether the model's performance for each racial subgroup is similar with and without race included.

The PRM tool will be applied to all the children that are born in the jurisdiction unless the family chooses to opt out of the project and does not want to receive any services from the providers. After using the PRM tool to score each child, the service provider will decide how to engage with the family and tailor services to their needs. Families with the highest risk will receive more intensive service offerings from the system. In contrast, the lower risk ones will receive universal services such as access to a support hotline or website. From the service provider point of view, the purpose of the tool is to reach and engage more families who might benefit from the support and connect families to services before incidents of abuse and neglect occur. Moreover, the goal is to ensure that families with the most complex needs receive prioritized access to the most intensive support.

According to the preventive programme's designers, it is a "universal graduated program", which means everyone has access to the universal child protection services provided by the programme, but only families with moderate to complex needs will receive more intensive service offerings. Based on their estimated risk score, a family will be allocated to the Priority tier, Family Support tier or Universal tier. Universal outreach is offered to any mother who gives birth in the jurisdiction regardless of need. Meanwhile, the Family Support tier is designed for only families with moderate risks/needs and includes more support, such as Home Visiting programmes. The Priority tier is dedicated to the highest risk group to provide them with the most intensive services.

The preventive PRM tool was trained on a range of adverse outcomes that can be generated using the available data. Specifically, the model designers tested various dependent variables (e.g., child welfare referrals, case openings, out-of-home

placements/removals, out-of-home placement due to physical abuse only) with different follow-up periods (e g., one year, two years, three years). According to the modellers, predicting adverse events in a child's early years is critical for intervention since the early experience of adversities will have a long-term impact on the child's development (Felitti et al., 1998).

After training the model on various outcomes, the tool designers decided to use out-of-home placement by the child's third birthday as the final target outcome. By predicting out-of-home placement in the first few years of a child's life, the modellers argue that they focus on a fundamental period in childhood development. They explained that early exposure to adversities and maltreatment could prolong the consequences and impede the child's ability to develop fully. Moreover, they achieved higher predictive accuracy when training the model on foster care placement. They also asserted that out-of-home removal demonstrates child protection involvement which can be prevented by early intervention.

However, there is concern regarding the training on foster care placement as it might not be a good indicator of actual maltreatment. In other words, out-of-home removal might not truly reflect the underlying maltreatment risk that the child is facing: "if the child protection system is poor at identifying children at risk of harm, then training a model on children who are removed might not actually identify children at true risk of harm" (Vaithianathan et al., 2020, p.18) Ideally, families eligible for Priority tier services should be at high risk of a range of (objective) adverse outcomes, including infant mortality, deaths associated with maltreatment and actual severe abuse and neglect. Training the model on such outcomes could also be problematic. For instance, if the algorithm designers were to use mortality as the outcome variable, they face the unbalanced dataset issue. That is, one class label's prevalence rate (e.g., the class label indicating mortality) is significantly lower than that of the other class (e.g., the class label indicating being alive). This is also true of placement but the issue is much less severe. The placement base rate is roughly 18 removals per 1000 children, whereas the mortality rate is approximate 6 deaths per 1000 children. Fortunately, there were very few children who died between 2012 and 2015 in the jurisdiction. Additionally, the modellers could not train the PRM tool on abuse and neglect as the data on these outcomes are not readily available. Therefore, they built the preventive model on home removals after taking into account all the above considerations.

To train the algorithm, the data set was split into two subsets according to the 70:30 rule for training and testing. The tool designers used the LASSO method to estimate the likelihood of foster care home placement by age three. LASSO is a Regularized Logistic

Regression algorithm (Tibshirani, 1996). The technique is also called regularised regression that adds constraints to the loss function to ensure that the model predicts better on unseen data.

Let y be the dependent variable, or 'out-of-home placement by age three' outcome. X is an independent variable $N \times p$ matrix with p = 459 input variables and N = 52,520 children. The LASSO model minimises the log loss function subject to a constraint:

$$\min_{\beta_0,\beta} -\sum_{i=1}^{N}\left[-ln\left(1+e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}\right)+y_i(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})\right]+\lambda\sum_{j=1}^{p}|\beta_j|$$
$$= \underbrace{L_{log}}_{Log\ loss} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{Regularization\ terms} \quad (1)$$

Like a logistic regression, the LASSO algorithm estimates coefficients that best fit the data by minimizing the function above. However, unlike logistic regression, which only minimizes the log loss function (i.e., L_{log}) (Hastie et al., 2021), LASSO needs to minimize the summation of log loss and the regularisation term (called the "total loss"). By adding the second term $\lambda\sum_{j=1}^{p}|\beta_j|$, which is also called the shrinkage penalty, LASSO will force some of the estimated coefficients to zero to minimize the total loss. As a result, the number of weighted predictors in the LASSO model will be smaller than the number of estimated coefficients produced by an ordinary logistic regression. The shrinkage term in the LASSO model thus acts as the variable selector which keeps only most important variables in the final model. As a result, the LASSO model becomes less flexible than a logistic estimator due to its more simplified form with fewer non-zero coefficients. This makes the LASSO algorithm avoid overfitting, which refers to a model that is very good at fitting its training data but not as good when predicting new data. However, the shrinkage penalty will also distort the values of the other betas. This means that the estimated LASSO probabilities are biased estimates of the "true" risk.

In machine learning, the trade-off between variance and bias is important. Variance tells us how much the estimated model will change when we change the training data. Normally, if the functional form of the estimated model is so complex and flexible that it precisely predicts all the data points in the training data set (i.e., by drawing a curve that fits through all the data points), small changes in the training data will lead to significant change in the estimator. In this case, the model is said to have high variance. In other words, the model follows all the noises in the training data too closely, which results in unnecessary learning of useless patterns. What it learns from the training set may not

be representative of the population. When applying a high variance algorithm on future data, its accuracy decreases significantly. In other words, a flexible model is more likely to suffer from the overfitting issue than a simple model.

On the other hand, bias refers to in-sample error (Kleinberg et al., 2015). Bias measures the total expected difference between estimated values and actual values of dependent variables in the training data. Bias occurs when we try to approximate real-life data by a simple model. For instance, if we estimate the relationship between Y and X using a simple method such as linear regression when the relationship is non-linear, we are undoubtedly going to get a biased prediction (James et al., 2013). In this case, a more complex functional form will result in less bias in prediction for the training data. However, by increasing the complexity of the model, we potentially increase its variance, which makes the algorithm less accurate at predicting future data. For prediction purposes, a low variance but high bias model may be more desirable than a low bias but high variance one (Kleinberg et al., 2015). In other words, in the case of prediction problems, we care more about how well the model predicts for future data than how it performs on the current training data.

By introducing the regularizer, $\sum_{j=1}^{p}|\beta_j|$, the LASSO model allows for high in-sample error while penalising high variance. Lambda determines the trade-off between variance and bias. Thus, the LASSO method is a biased estimator, but it can generalize well to unseen data due to its low variance compared to logistic regression. Since the purpose of the PRM tool is to accurately predict future children who will be at high risk of severe abuse and neglect, overfitting and high variance are of more concern than in-sample bias. Therefore, the LASSO model is favoured over ordinary logistic regression.

The LASSO model's estimated probabilities from the training data are further stratified into twenty risk scores with an approximately equal number of children in each score bin. That is, each score bin contains roughly 5% of the population. To do this, the modellers first sort the predicted probabilities in ascending order. Next, they divide the probability distribution range into twenty continuous intervals with equal proportions of the sample in each interval. Stratifying predicted probabilities into twenty risk score bins provides an easy way to distinguish children. Those who score 20 have much higher predicted probabilities of foster care placement by the third birthday than children who score 1. More importantly, risk scores produced by the tool will be used to prioritize children to different services. Moreover, due to limited resources allocated to the priority group, policymakers have decided that the Priority tier can only accommodate 5% of children. Therefore, children whose score is 20 will be offered the Priority tier, which is the most intensive service designed to support families with the most complex needs. Children

with scores from 17 to 19 are provided with less intensive Family Support services. Any children who score lower than 17 are considered low risk of future maltreatment; thus, they only receive Universal services.

The PRM tool's accuracy is assessed on the test set using multiple metrics. The reason for evaluating the model performance on the test set is because we care about how well the model predicts for new data. Note that the test set includes data on the 3-year placement outcome of the children in the sample, so predictions can be compared to actual outcomes. The first metric is the true positive rate (TPR) which tells us the share of all removed children who were classified at birth as "high" risk (i.e., score 20). A high TPR means a large proportion of vulnerable new-borns are correctly identified by the tool as eligible for intensive service engagement. The second metric is positive predictive value (PPV) which tells us the percentage of placed children amongst those classified as "high" risk. A high PPV suggests that a significant number of eligible children experience placement by age three. It is important to note that there are no interventions for children in the test set. The TPR and PPV for eligible children in the test set are 54% and 20%, respectively.

Another general way to demonstrate how the estimated risk scores are predictive of placement outcome is using the ROC (Receiver Operator Characteristic) curve. A ROC curve shows the performance of the PRM tool at any high-risk threshold, α. There are two parameters in this curve. *Sensitivity*, or the TPR, assesses the algorithm's ability to correctly identify children who will be placed by age three. In contrast, *Specificity* is the proportion of non-placed children who are correctly classified as low risk: the true negative rate (TNR). The false positive rate (FPR – the proportion of non-placed children who are mis-predicted as high risk of placement) is 1-TNR. When varying the high-risk threshold, α, we receive different combinations of TPR and FPR and the ROC curve plots these combinations. Because TPR and FPR both increase with the threshold there is a trade-off. In other words, we cannot increase TPR while also decreasing FPR. Ideally, we would want to choose a decision cut-off that gives the highest possible TPR and lowest FPR: a point on the ROC curve that is as far as possible from the diagonal. However, in the case of the PRM tool we are studying, the service providers can only serve 5% of the population due to limited resources. This constrains the threshold.

The area under the ROC curve (AUC) provides a generalised measure of model performance across all possible high-risk cut-offs. It shows the algorithm's ability to accurately rank a randomly placed child more highly than a random non-placed child in terms of predicted probabilities (Hanley & McNeil, 1982). AUC values range from 0 to 1. A PRM that is perfectly able to separate placed children from non-placed children has

an AUC of 1. Conversely, a PRM with an AUC of 0.5 performs no better than tossing a coin. ROC curves and the AUC are analogous to Lorenz curves and Gini coefficients (respectively) for describing income inequality. The AUC summarises the degree of outcome-risk-inequality across the risk scores: in this case, more inequality is better (i.e., very low risk at one extreme and very high risk at the other). The PRM tool that we study here has an AUC of 92.4% on the test set which shows that the model is highly discriminating of out-of-home placement.

The tool trained on placement is also externally validated using various other adverse outcome measures. This is to prove that the PRM is not only predictive of its trained outcome (i.e., placement), but it is also good at identifying other adversities that the families and the new-borns might experience. A set of external outcomes, including maternal homelessness, maternal jail stays, maternal mortality, child mortality (any causes), post-neonatal fatality and violent, accidental, maltreatment-related mortality/near mortality (sensitive death), was constructed to externally validate the model. The model validation process includes calculating the PPV at each risk score level using each validated outcome and comparing the calculated PPV values of each outcome between different risk score bins. If the model is good at predicting a particular external outcome, children who score higher will have a higher likelihood of having that outcome compared to children who score lower. The model developers show that children and families eligible for the most intensive services face higher risk of each of these adversities compared to non-eligible children. In particular they found that children who score 20 are 30 times more likely to have mothers who served jail time and 27 times more likely to have mothers who died, than children who score lower. Those children are also 3.5 times more likely to die from any cause than other children.

The PRM tool offers a few advantages over traditional actuarial assessment tools. One advantage of using PRM tools is that it does not require participants to provide additional information by answering lengthy and invasive questionnaires. Instead, the algorithm can draw upon verifiable data that are far more complex and detailed, and accurate than self-reported information from clients. PRMs are also capable of screening a whole population of interest and in particular, they are automated and can be scaled up. Moreover, the PRM tool is capable of proactively identify high-risk cases even when there has been no previous service contact. Many families who are afraid of stigmatization when using the services might not want to engage with child welfare services and are therefore not known to the service providers. Thus, the preventive PRM can reduce information barriers and unwarranted selection of easily engaged children whose risk is low.

On the other hand, there is growing concern that the use of PRM may potentially reinforce any racial bias prevailing in the CPS system (Gillingham, 2016; Glaberson, 2019) when these tools are trained on outcomes in the child welfare (such as removals). There is mounting evidence that the system is biased against children of colour (Cénat et al., 2021; Rivaux et al., 2008), and because the recorded data is inevitably a reflection of history, people are worried that any such bias can be "hardwired" into the algorithm; thus, it may make decisions that discriminate against black children. However, the evidence is mixed as to whether the tools alleviate or exacerbate racial bias (Drake et al., 2020). In particular, the PRM tool used to make screening decisions by child welfare staff in Allegheny County, Pennsylvania (Allegheny Family Screening Tool - AFST), has been shown to reduce disparities in case opening rates across racial groups (Goldhaber-Fiebert & Prince, 2019). However, just because we did not see racial bias in the AFST model, this does not guarantee that other PRM tools are also free of bias. Thus, careful design and thorough examination of racial discrimination in the PRM tool are always necessary to understand whether the model helps eliminate any existing bias (Gillingham, 2016; Glaberson, 2019).

According to the tool designers, the PRM tool is relatively predictive of out-of-home placement by age three for the whole population, however; it has differential/disparate accuracy by race. By calculating the PPV for eligible black and white children in the Priority tier, the tool's designers found that the model is much more accurate at identifying high-risk white children who are removed from their parents' care than high-risk black children. In this dissertation, we will discuss further the underlying disparity in out-of-home placement predictions by race. More importantly, we find that the use of a proxy variable, which is foster care placement in our context, can potentially lead to greater racial disparity in child maltreatment predictions - in particular, in child mortality caused by abuse and neglect ("sensitive death"). *Figure 1* summarises our finding on the racial differences between estimated true risk of placement versus true risk of sensitive death. The subsequent sections analyse this phenomenon, its causes and implications, in more detail.

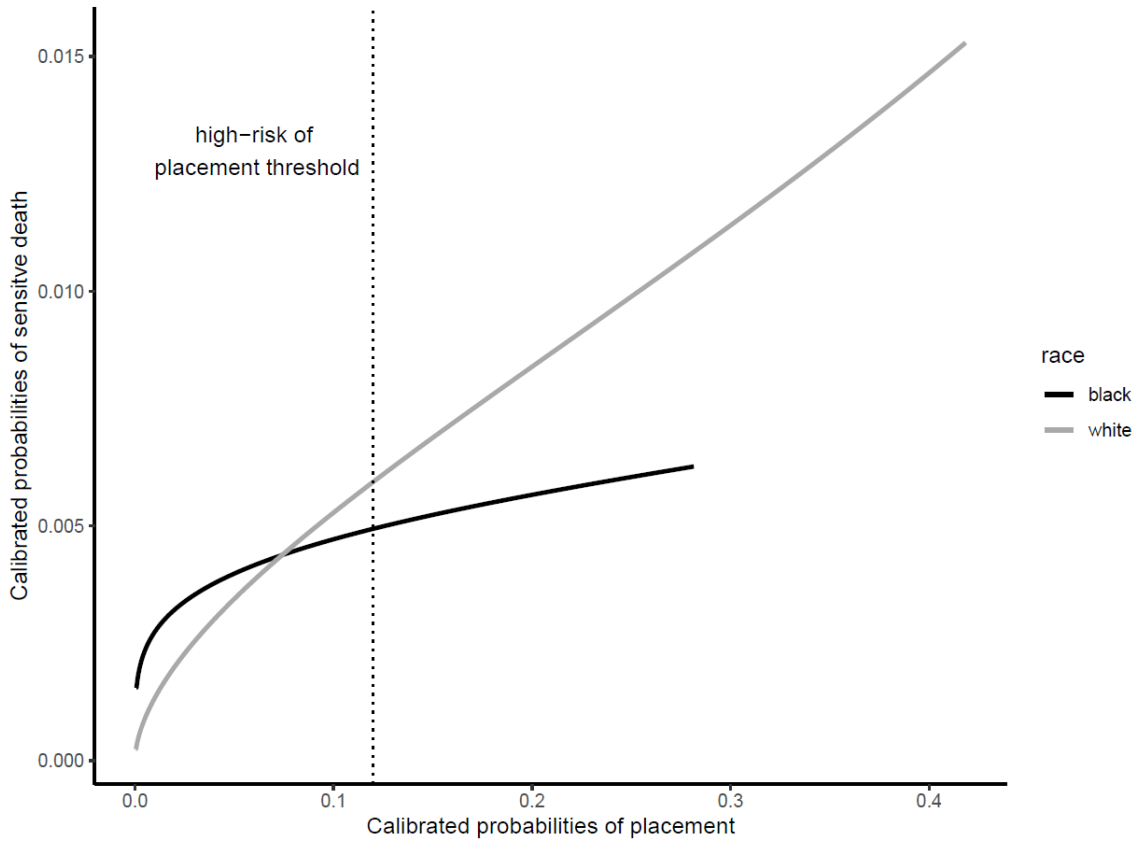

Figure 1: Actual placement risk versus sensitive death risk, by race.

In this paper we discuss the following three things. (1) First, although the literature pays a lot of attention to error rate balance conditions (i.e., equal FNR and FPR across race groups) when assessing whether an algorithm is "fair" (a term, which we will define more precisely in a later section), we argue that this is misguided. We prove that even if the PRM tool is equally accurate in predicting the outcome for different racial groups (i.e., perfectly "fair"), it still fails to meet the error rate balance (ERB) condition due to the differences in the underlying risk distribution across races. That is, when the actual risk distributions are dissimilar by race, ERB is necessarily violated. (2) Second, we argue that there is mis-ranking in the tool's predictions of children's true risk, such that black children eligible for the Priority service tier are at lower risk of placement and abuse than white children who are prioritised for the same service. (3) Finally, as we have shown in *Figure 1,* while the placement risk is a good proxy for abuse risk for black and white children (both curves in Fig 1 are upward sloping), the strength of the correlation varies by race and level of risk (*Fig 1*). More specifically, out-of-home placement is a better proxy for abuse for white than for black children.

2.2. Outline of the dissertation

The remainder of this dissertation is structured as follows. In Section 3 we will explain the notion of define algorithmic bias as well as discuss some of its common causes. In particular, we will focus on important factors leading to algorithmic bias, which is mismeasurement by training on a proxy outcome, and also provide an exemplar of this issue. We then pose our research questions in Section 4. In Section 5 we define a set of standard fairness metrics commonly cited in the algorithmic bias literature. In Section 6, we analyse algorithmic fairness of the studied PRM tool in terms of predicting for placement and child maltreatment outcomes. We then provide a possible explanation as to why we observe disparities in placement and child maltreatment predictions across race (Section 7). Finally, we conclude and discuss some limitations in Section 8. All the proofs and additional tables are placed in the Appendix.

2.3. Data setup

The empirical results in this dissertation are based on a subset of the data used to train the studied PRM tool. The author obtained ethics approval to access the data[1]. The data consists of the PRM ventile risk scores, LASSO predicted probabilities, the dependent variable (i.e., indicator for out-of-home placement by age three), sensitive death outcome (i.e., indicator for death caused by violence, accident, child maltreatment), and race variables. We mainly concentrate on studying the difference between black (b) and white (w) children; thus, we will filter the original data to the subset comprising these two groups. We use race variables defined by the model owner to classify children into black and white groups. After excluding other races (i.e., Asian, Hispanic, Native American), the new data set contains 48,034 children (N = 48,034)[2], where African-American children account for one-fourth of the total number of children ($n_b = 10{,}367$) and the rest are white ($n_w = 37{,}667$).

[1] Ethics application 20/245: Understanding the mechanism of racial bias in predictive risk models of child welfare. See Appendix I for the approved letter.

[2]The PRM tool was built off the set of 52,520 children. However, for the purpose of our analysis, we limit our attention to only black and white children. Therefore, the data set used in the later analysis consists of 48,034 children.

3 What is Algorithmic bias?

3.1. Algorithmic bias and its causes

Algorithms like the PRM tool have recently been embedded in high-stakes decision-making processes in child welfare systems across nations (Drake et al., 2020; Gillingham, 2016; Glaberson, 2019). Views on the use of these tools are contentious. Some believe that the algorithms will help fix the flaws in current practice by accurately triaging at-risk children and reducing racial bias (Cuccaro-Alamin et al., 2017; Drake et al., 2020; Goldhaber-Fiebert & Prince, 2019). Conversely, many hesitate to embrace the new technology with particular concern over the fairness problem. It is a very natural and human tendency to be cautious of the unknown. Many of these algorithms are essentially a "black-box". Therefore, even the people who design and own the tool sometimes do not fully understand what happens inside the box.

In the child welfare context, critics have focused on how the system treats minority children differently from children belonging to the majority class (Chibnall et al., 2003). The introduction of the PRM tool, which is built using child welfare data, certainly invites more criticism. In particular people argue that the data is a reflection of the system history; thus, the past dwells within the algorithm and necessarily exacerbates the existing bias in the predictions (Glaberson, 2019). In recent years, researchers are increasingly interested in understanding algorithmic fairness (bias) in various applications, including the child protective field. However, the study of machine learning fairness remains immature (Chouldechova & Roth, 2018).

In this section, we will introduce the concept of algorithmic bias and its leading causes. To end this section, we also motivate our research by providing an exemplar of machine bias in a health care prioritization tool used in the US.

Algorithmic bias generally refers to the situation where the model prediction is not as accurate for one group as another. When discussing fairness, it is important to mention two critical notions: disparate impact and disparate treatment. Treating people differently based on their group membership is defined as disparate treatment (direct discrimination). On the other hand, disparate impact means "negatively affecting members of a protected class more than others even if by a seemingly neutral policy (indirect discrimination)" (Pessach & Shmueli, 2020). Algorithms can avoid direct discrimination by giving different people the same treatment (i.e., a universal risk threshold is systematically applied for everyone).

Even when the first type of discrimination can be eliminated, people are still sceptical about disparate impact. If the algorithm is essentially less accurate at predicting the

outcome for the minority group, this could significantly disadvantage people belonging to this group. According to Chouldechova (2017), black defendants, whether they recidivate or not, on average, receive more severe penalties than their white counterparts when using a biased algorithm such as COMPAS tool, which was developed to grant bail or parole for defendants.

However, one could argue that the PRM tool that we are concerned with in the present study does not serve a punitive purpose like the algorithm used in the criminal justice system. We learned from the previous section that the PRM tool supports the service prioritization process, and participation in all the services is voluntary. In this case, even when the tool is less accurate for black families -- it mistakenly offers black families more intensive services -- black families will not suffer from this misclassification as they always have an opt-out option to consider. However, if they take up the benefits even when they don't need them, the service providers will miss out on opportunities to serve in-need families. The PRM tool designers argue that the disparate impact is less likely to disadvantage black children. Nonetheless, it is still costly for society, with potential harm to unprotected-children.

There are several common reasons for algorithmic bias to exist. Firstly, bias can stem from the data used to train the algorithm. Researchers often use the expression "bias in, bias out" to summarise this type of bias (Barocas & Selbst, 2016; Glaberson, 2019). In particular, when the training data already includes human prejudice or stereotyping, learning well from these examples will naturally replicate the same dynamics (Chouldechova & Roth, 2018). Moreover, training algorithms on incomplete or unrepresentative data to generalize for the whole population may lead to biased predictions (Cowgill & Tucker, 2019; Pessach & Shmueli, 2020).

The second source of bias originates from the objective of the training methods, which usually attempt to minimize the overall error. While the algorithm is trained to minimize the total error, if it cannot simultaneously fit both the minority and majority groups optimally, it generally fits the majority better since the majority population contributes more to the total error (Chouldechova & Roth, 2018; Pessach & Shmueli, 2020).

Measurement error is recently emerging as an understudied source of bias in the algorithmic fairness literature (Kleinberg et al., 2015; Mullainathan & Obermeyer, 2017, 2021). In particular, the choice of a biased proxy as the model's target variable could induce algorithmic bias. Measurement error is defined as the difference (error or Δ) between the measured outcome (Y) and the outcome of interest (Y^*) (e.g., Y = placement and Y* = abuse in the context of PRM algorithm):

$$Y = Y^* + \underbrace{\Delta}_{measurement\ error}$$

Defining target variables is somewhat tricky as it requires turning an abstract idea into a measurable subject (Passi & Barocas, 2019). Naturally, the tool's developers will choose proxy variables that are readily available to them, assuming that this measured outcome is the best proxy variable for the ground truth. For instance, in the child protective system (CPS), out-of-home placement is a proxy for child maltreatment since they are highly correlated; however, out-of-home placement is not the same as child maltreatment. There are many children who continue to be abused and are never discovered and there are, conversely, children who are removed who were not at risk of abuse or neglect.

Unlike the causal inference task, we are not interested in knowing if delta is correlated with any of the predictors in the model for the prediction task (Mullainathan & Obermeyer, 2017). In the causal inference task, we want to make sure our estimators are unbiased. However, the prediction task only requires low error in predicting the proxy outcome. If the error is a white noise that is entirely uncorrelated with input variables (X), on average, a good estimator $\hat{y}$ will approximate the predictable part of the ground truth.

$$\hat{y} \approx E[Y|X] = \underbrace{E[Y^*|X]}_{signal} + \underbrace{\overbrace{E[\Delta|X]}^{\to 0}}_{White\ noise}$$ (assuming expected value of prediction error is 0)

On the other hand, if the error part is not "nicely behaved" (i.e., white noise), estimated predictions will depend on both signals and predictable error. When measurement error becomes more predictable than the underlying risk (i.e., signal), it will dominate predictions and might distort the final decisions (Mullainathan & Obermeyer, 2021). Thus, the relative predictability of the ground truth and the measurement error will determine how severe the bias is.

$$\hat{y} \approx E[Y|X] = \underbrace{E[Y^*|X]}_{signal} + \underbrace{\overbrace{E[\Delta|X]}^{\neq 0}}_{Predictable\ error}$$

However, predictability of error will not affect decision-making when it does not lead to a risk ranking distortion. Suppose child welfare workers overestimate the maltreatment risk for all children; thus, their estimated risks now move up by an equal amount. However, this error caused by CPS workers does not impact the rank ordering of children in terms of their actual risk. In this scenario, the tool is still accurate in classifying children according to their true risk.

Conversely, when mismeasurement in target variables distorts the ranking of actual risk, this could cause serious problems. For instance, due to racial bias in the CPS system, black children are removed from home more often than otherwise similar white children.

Although both black and white children might be abused and neglected at the same rate, the model trained on out-of-home placement, a proxy for actual maltreatment, will rank black children as at higher risk than similarly risk-exposed white children. In other words, the rank ordering of out-of-home placement is not the same as the ranking order of the underlying risk. In this case, decisions and allocations produced by the tool will be biased.

Since proxy choice directly affects algorithmic bias, careful selection of proxy variables will bring about significant benefits. Next, we will take a closer look at how measurement error plays out in one particular use case, which is the health prioritization algorithm used in the US health care system.

3.2. Obermeyer et al., (2019)

In light of the research on mismeasurement, Obermeyer et al., (2019) conducted a study on how the choice of the proxy outcome can potentially lead to racial disparity in terms of future predicted health outcomes. The authors investigate racial bias in a risk-prediction tool widely used by the US healthcare system. Health care professionals and funders rely on these tools to allocate more intensive services to the most high-risk patients as predicted by the algorithm. Ultimately, the purpose of these "high-risk care programmes" is to improve the health outcomes of complex-need patients by additional service provision. Since these health care programmes are resource-intensive and costly, the tools need to correctly identify patients who will benefit the most from them. However, triaging the patients who genuinely have complex health needs and will derive the most significant benefit from these services is a big challenge. It is a causal inference problem that asks whether these intensive treatments do in fact improve the sickest patients' health outcomes. To measure the counterfactual - what happens with or without these programmes - requires estimation of treatment effects at the individual level. Put differently, a perfect risk score (S^*) could have the form: $S^* = E[H_{1i} - H_{0i}|X_i]$ (where $H_1 \; and \; H_0$ are treated and untreated health outcomes, respectively) (Mullainathan & Obermeyer, 2021). But estimating the expectation on the right-hand side (RHS) is a very daunting task. To overcome this issue, the algorithms' designers used healthcare costs to predict future health care needs. A fundamental assumption is drawn: the sickest patients are also the costliest; thus, they are the patients with the greatest care needs who might benefit the most from the programme. By making that assumption, the tool's developers hope that the algorithm, which is strongly predictive of total health expenditure, will also be predictive of health care needs, for which it was not trained. Although, health care cost and health care needs are strongly correlated, they are not the same.

The authors employ extensive hospital data consisting of all primary care patients from 2013 to 2015. There are 6,079 Black patients and 43,539 White patients in the data set. Typically, patients with risk scores of 97th percentile or above will be automatically enrolled in the care management services. Patients whose scores rank above the 55th percentile will be referred to their doctors to consider whether to continue referring them to the programmes based on these patients' available information. According to Obermeyer et al. (2019), there are various fairness metrics that can be used to assess algorithmic bias[3].

These fairness concepts will be discussed in more detail in a later section. In this section we only introduce them briefly. Specifically, the fairness metrics are calibration (i.e., people who are predicted at the same risk score should have equal chance of having an outcome), predictive parity (i.e., predicted high-risk individuals have the same probabilities of a future event regardless of their race group membership), statistical parity (i.e., the percentage of people being assessed as high risk is equal across groups), and error rate balance (i.e., the percentage of people the model mis-predicts their risks with respect to their true outcome, is the same across groups). Statistical parity seems inappropriate in the context of the health care tool as it is undesirable to have the same proportion of black and white patients in the 97th percentile group regardless of their real needs. Obermeyer et al., (2019) states that they concentrate on calibration criterion as it is most relevant to the real-world application of the tool. As explained by the authors, the tool's purpose is to correctly capture as many patients as possible in the eligible group with complex needs, in order to provide them with early treatment intervention. In other words, the authors wanted to know whether the tool is equally well at predicting the complex-health needs for black and white patients. In particular, the authors check whether the tool is equally well-calibrated across racial groups regarding health outcomes and health care expenditure. That is, black and white patients who are eligible for intensive care programmes should be equally sick, and they should also generate the same amount of healthcare funding. The authors use the number of active chronic conditions as a proxy for patients' health outcomes, as it is a good indicator of a patient's health. Other biomarkers such as high blood pressure, the severity of diabetes, high cholesterol, renal failure, and anaemia are also employed to measure the complexity and severity of patients' health needs.

The authors find that the risk assessment tool is well-calibrated for its trained outcome - total medical expenditure. At any given risk score, black and white patients generate the

[3] Throughout this dissertation, we will continue to refer to algorithmic bias as the PRM tool's ability to predict one particular group better than the others. In other words, if a model exhibits algorithmic bias, it is racially biased against one particular group in terms of its predictions. On the other hand, when referring to measurement bias, we refer to one of the causes of algorithmic bias as explained above.

same health costs. However, conditional on risk scores, black patients are significantly sicker than white patients anywhere in the risk distribution. Among the patients who score at or above the 97th percentile, which is also the auto-enrolled threshold, eligible white patients have 26.3% fewer chronic conditions than black patients. Also, African-American patients suffer more from diabetes, renal failure, severe hypertension and high cholesterol than their white counterparts. The authors attempt to quantify the substantial disparities in health status across racial groups. They replaced healthier white patients whose risk score was just above a fixed threshold, with sicker black patients whose risk score was just under that same threshold. By continuing the replacement process until the marginal patient is equally healthy, the proportion of black patients above the 97th percentile increased from 17.7% to 46.5%.

Common sense would suggest that less healthy patients need more health care, hence generating more health expenditure than healthier patients. However, according to the authors' reasoning, there are a few channels which help explain why sicker black patients pay less for their health care than their white counterparts. To see this, the authors compare total medical costs against health outcome. Conditional on the number of chronic conditions, black patients generate lower medical costs than white patients. The authors assert that "accurate prediction of costs necessarily means being racially biased on health" (ibid., p.4). Multiple factors affect the relationship between health outcome and heath care expenditure. Although in this study, Obermeyer et at. (2019) restrict attention to insured patients, and explains that poverty can affect health care utilisation through multiple channels. In particular, poor patients have a lower ability and less inclination to access health care services than wealthy patients when sick due to barriers such as "geography and differential access to transportation, competing demands from jobs or childcare, or knowledge of reasons to seek care" (Obermeyer et al., 2019, p. 4). For instance, a poor black patient who lives remotely from the registered hospital and works two low-paid jobs is less likely to have regular check-ups. In contrast, a wealthy white patient who lives near a hospital and works a high-paid job is more likely to visit their physician regularly. Thus, even though they might be equally unhealthy, the wealthy white patient will generate more medical costs than the poor black patient.

The authors also seek to find other explanations as to why race could affect health care utilization. Black patients are more willing to take up recommended preventive services when their physicians are black. Furthermore, black patients have a lower level of trust in the health care system than white patients. So, they are less likely to listen to their doctors' advice. Besides, healthcare staff also perceive black patients differently from white patients regarding knowledge or pain tolerance. As a result, black and white patients who are equally sick might receive different treatment paths due to doctor bias

or patient preference. Combining all these factors, black patients possibly receive significantly lower health expenditure than white patients. Thus, by using these tools to predict complex future health needs, one may be likely to provide more health resources to white patients who have similar or fewer health conditions than black patients.

Obermeyer et al. (2019) conclude that bias can arise due to a proxy variable's choice. Total medical cost seems to be a reasonable choice as patients with the greatest future health cost could also most benefit from the care management programme. However, as medical expenditure is subject to racial bias using it as the proxy for health needs is likely to disadvantage black patients. The authors suggest using the number of active chronic health conditions or avoidable future cost due to emergency visits or hospitalizations as outcome variables to build the risk-assessment tool for all patients. Obermeyer and his colleagues found similar model performance when training the algorithm on the number of active chronic health conditions and total medical costs. Nonetheless, they also found less racial bias in the top highest-risk patients when using the number of active chronic health conditions as a proxy for health care needs. The study emphasizes the importance of proxy variable choice in building prediction algorithms such that "careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their risks" (p. 7).

4. Research Questions

1. Does the PRM tool used in child welfare systems exhibit racial disparity in out-of-home placement predictions?
2. Does the use of a proxy variable, which is foster care placement in our context, potentially lead to racial disparity in child maltreatment predictions?

In this section, we thoroughly examine whether the PRM tool, which was developed to be used in conjunction with the child maltreatment prevention programme in the US child protective service (CPS) system, is racially biased against a particular group when predicting home-removal outcomes. As explained previously, the PRM tool can provide a means to better engage families who might benefit from the support, and connect families to services before incidents of abuse and neglect occur. Thus, the tool needs to ensure accurate identification of the families with the most complex needs to be prioritised to the most intensive support. In addition, so long as the families and the children face the same level of adversity, their needs should be estimated equally, regardless of their race. Therefore, assessing whether the tool precisely predicts the level of risk for families across races is essential. More importantly, racial disparities in the child welfare system is also a hot topic that attracts attention and criticism from communities, lawmakers and researchers. Given that this algorithm may have an impact on many lives and aid decision-making regarding the safety of many children, investigating the racial discrimination aspect to see whether the model helps eliminate any existing racial bias is worthy of investigation.

The PRM tool is built using out-of-home placement data, which is a proxy for abuse and neglect. However, there is concern regarding the training data on foster care placement as it might not be a good indicator of actual maltreatment. In other words, out-of-home removal might not truly reflect the underlying maltreatment risk that the children are facing (Vaithianathan et al., 2020). Furthermore, we also learned from the previous section how proxy variables could lead to algorithmic bias. Therefore, we argue that it is vital to understand if out-of-home placement can cause racial discrimination in child maltreatment predictions. In the rest of this dissertation, we will address the two questions in more detail.

5. Definitions of algorithmic fairness

In order to determine whether the PRM is racially biased against one particular group, the current dissertation will examine a standard set of fairness metrics. These metrics are calibration, predictive parity, error rate balance and statistical parity. Chouldechova (2017) summarises several fairness criteria that have been used in the recent literature. A tool that satisfies calibration across racial groups should equally well predict the likelihood of a future event for all that share the same predictive risk score, regardless of their group membership. To define calibration formally, let S be the score variable, let R denote race, either R=b (black) or R=w (white), and let Y be the indicator, with Y=1 if the person has the outcome (e.g., the defendant commits new felonies upon release), and Y=0 if otherwise. Calibration (across groups) requires:

$$Pr(Y=1|S=s, R=b) = Pr(Y=1|S=s, R=w) \quad (1)$$

The second fairness notion is predictive parity, or equal PPV (positive predictive value). Given a high-risk cut-off score (s_{HR}), which means that everyone who scores above the cut-off is classified as high risk and everyone at or below s_{HR} as low risk, PPV measures the probability of the future event (in this case, re-offending) amongst those classified as high-risk. For example, if risk scores run from 1 to 10, then high-risk threshold values are from 0 to 9. In particular, if $s_{HR} = 0$ then all the people in the data set will be classified as high-risk. On the other hand, when s_{HR} = 9, only people who score 10 will be considered as high-risk.

Predictive parity is satisfied if high-risk individuals have the same probabilities of the future event regardless of their race - they have equal PPVs. While the calibration criterion is assessed at a single score level, predictive parity (PP) is calculated for the group of risk scores that is greater than the chosen threshold:

$$PPV_b = Pr(Y=1|S>s_{HR}, R= b) = Pr(Y=1|S>s_{HR},R = w) = PPV_w \quad (2)$$

The third fairness metric is error rate balance (ERB), which compares false positive rates (i.e., the proportion of instances with Y=0 who are misclassified as high-risk) and false negative rates (i.e., the proportion of instances with Y=1 who are misclassified as low-risk) across racial groups. A tool is said to meet error rate balance if it has equal false positive and false negative rates across racial groups.

$$FPR_b = P(S>s_{HR}|Y=0, R=b) = P(S>s_{HR}|Y=0, R=w)=FPR_w \quad (3)$$

$$FNR_b = P(S\leq s_{HR}|Y=1, R=b) = P(S\leq s_{HR}|Y=1, R=w)=FNR_w \quad (4)$$

The final fairness metric is statistical parity:

$$P(S>s_{HR}|R=b) = P(S>s_{HR}|R=w) \quad (5)$$

At a chosen high-risk cut-off, it requires the percentage of people being assessed as high risk to be equal across groups. According to Chouldechova (2017) statistical parity is much more relevant in contexts such as employment or college admission. It might be required by law that companies or colleges should employ or admit people from different racial or gender groups in the same proportion. However, in many other contexts, such as predicting recidivism or health care needs or child maltreatment, it is undesirable to try to achieve the same percentage of black and white in the high-risk category.

Chouldechova (2017) studies the fairness of a well-known recidivism prediction instrument (RPI) widely used by law enforcement and courts in the US. The tool is called COMPAS, which has been criticized by a ProPublica team, accusing it of racial bias against black offenders. Tools like COMPAS are risk assessment tools that use current demographic information and history to predict a future event's likelihood. In the case of COMPAS, it assesses how likely the offender will be to re-offend in the future. COMPAS scores are used in conjunction with judges' own experience to assess the possibility of committing new offences while on release or failing to appear on the trial day. Thus, this prediction tool aids judges to decide whether to release or detain criminal defendants before their trials. The increasing employment of these algorithms into high-stake decision-making processes attracts more and more attention, and criticism, from researchers, journalists, policymakers, and society. Thus, it is crucial to ensure that these instruments are accurate and fair to avoid inequitable outcomes for different racial groups.

By using Broward County's data, made public by the investigator team from ProPublica, Chouldechova re-evaluated the COMPAS tool's fairness using the three metrics (calibration, PP and ERB) defined above. The data set consists of risk scores scaled from one to ten, 2-year recidivism outcomes, and other demographic details as well as crime history variables for the period from 2013 to 2014. There are 6150 individuals in the data, of which 3696 are black, and the rest are non-black defendants. After re-evaluating all fairness criteria using different high-risk thresholds, the author finds that COMPAS is nearly well-calibrated. It also meets the predictive parity condition when choosing the high-risk threshold of 4; however, it always exhibits an error rate imbalance.

At first glance, calibration and predictive parity look relatively similar; however, a tool that satisfies the calibration criterion may still fail to meet predictive parity at a given threshold. In the case of COMPAS, the tool is well-calibrated across racial groups at any level of risk score from 1 to 10. Yet, it does not satisfy predictive parity when choosing a high-

risk threshold of 1. In other words, when aggregating the data into one big set containing all the defendants who score greater than 1, the tool fails to meet predictive parity. This phenomenon is related to Simpson's Paradox. It occurs when trends hold for separated groups but disappear or reverse once the data is aggregated.

According to Chouldechova (2017), predictive parity and error rate balance cannot be satisfied at the same time. She shows, through a compact equation, the impossibility of simultaneously achieving predictive parity and error rate balance without an equal prevalence rate (i.e., the percentage of people who have the outcome (Y=1)) across racial groups. She writes: "when the recidivism prevalence - that is, the base rate P(Y=1|R=r) - differs across groups, any instrument that satisfies predictive parity at a given threshold s_{HR} must have imbalanced false positive or false negative error rates at that threshold" (Chouldechova, 2017, p.157). In particular, the author asserts the following relationship between FPR, FNR, base rate (p) and PPV:

$$\text{FPR} = \left(\frac{p}{1-p}\right)\left(\frac{1-PPV}{PPV}\right)(1 - FNR) \tag{6}$$

This functional relationship was not proved in Chouldechova's paper, so we provide a proof in Appendix A.

From this equation, one can see that when predictive parity holds (i.e., PPV is equal across groups), the group with a higher recidivism rate (p) will also have a higher false-positive rate or a lower false-negative rate (Chouldechova, 2017). Because the predictive parity criteria calculated for the recidivism prediction instrument is met and black defendants have a much higher recidivism rate, their FPR will generally be higher, or their FNR will generally be lower, than that of white defendants, based on the equation (6). Indeed, this is consistent with what we observe in the COMPASS data. At any chosen level of cut-off, FPR is always higher for black defendants compared to white defendants, whereas the opposite direction was found when comparing FNR across race.

It is also essential to understand the extent to which dissimilar FPRs and FNRs can create disparate impact, meaning a stricter penalty for high-risk defendants of a particular racial group. To demonstrate the difference in outcomes for black defendants, Chouldechova (2017) considers a simple risk-based model to assign penalties as follows:

$$T = \begin{cases} t_{max} \; if \; s > s_{HR} \\ t_{min} \; if \; s \leq \; s_{HR} \end{cases} \tag{7}$$

where T is the penalty required from a defendant where t_{min}<T< $t_{\max}$. Naturally, high-risk defendants will receive a higher penalty than low-risk defendants. If the offender is predicted as high risk of recidivism, it represents the average penalty that high-risk

defendants receive. On the other hand, it represents the average remand time for defendants when they are predicted as low risk of recidivism. The author then quantifies the disparate impact by calculating the following quantity:

$$\Delta = \Delta(y_1, y_2) \equiv E(T|R = b, Y = y_1) - E(T|R = w, Y = y_2)$$

Delta is the expected difference in penalty between black and white defendants where $y_1, y_2 \in \{0,1\}$ are recidivism outcomes. The author then proves (ibid., Proposition 3.1) that

$$\Delta(0,0) = (t_{\max} - t_{\min}) (FPR_b - FPR_w)$$

and

$$\Delta(1,1) = (t_{\max} - t_{\min}) (FNR_w - FNR_b)$$

Recall that the observed differences between FPR and FNR for black and white defendants always persist in the tool, regardless of chosen cut-off. Thus, black defendants will receive heavier penalties (i.e., less remand time) than white defendants in both the non-recidivating and the recidivating subgroups. Thus, even when a tool satisfies predictive parity, it can still produce a disparate impact, penalising the racial group with a higher prevalence rate.

As explained by Chouldechova, her paper's empirical findings might be misleading if the observed outcome is subject to measurement error. It could be that the observed recidivism rates for both black and white groups do not truly reflect reality. Perhaps some individuals who are flagged as non-recidivists did indeed re-offend. However, the police could not identify their criminal activities, so they have no record of re-offence in the observed data. A large body of literature in the economic theory of discrimination suggests that police officers are taste-based discriminators against black defendants (Becker, 1957; Knowles et al., 2015). Taste-base discrimination refers to when the police officer acts as if they receive a higher payoff (or face a higher cost) when they search African-American pedestrians for carrying drugs (or let African-Americans walk past without stopping them to search). Thus, if black pedestrians are more likely to be stopped and searched, the arrest rate for this group may be higher.

On the other hand, when otherwise similar white people carry contraband, the police often do not search them, so their crimes are not recorded. Moreover, police officers are also more likely to patrol the area where black communities reside; thus, they are more likely to encounter criminal activities. Therefore, the empirical findings of fairness assessments might look different if accurate recidivism data were available. Furthermore, if the difference in the base rates of actual recidivism across racial groups

is trivial, the error rate balance might look less dissimilar. We, therefore, might expect to see that the false-positive rate for true recidivism is lower for black defendants than the current findings. Similarly, white defendants could have a lower false-negative rate.

To sum up, there are three standard criteria commonly used in the algorithmic fairness literature: calibration, predictive parity (PP), and error rate balance (ERB). Nonetheless, it is essential to note that when the prevalence rates are significantly different across races, PP and ERB cannot be met simultaneously. Last, but not least, when the ERB cannot be achieved, this will differentially affect the outcomes for each race. In the context of the COMPAS tool, although PP is satisfied, FPR is always higher, and FNR is lower, for black defendants regardless of chosen high-risk thresholds. Thus, if applying the policy as described above, black defendants always receive heavier penalties compared to white defendants, among both recidivists and non-recidivists. In other words, even when the tool satisfies PP criteria, it can still create disparate impacts due to the error rate imbalance issue.

6. Fairness analysis for the current PRM tool

The previous section explains the preventive PRM tool used in the US child welfare system, and we learned how the tool is constructed and used. More importantly, the tool developers concluded that more eligible white children are placed in foster care homes by their third birthday than similar black children. However, the algorithm designers did not explain the reason for the observed disparities. Motivated by this finding, we will focus on understanding the difference (by race) in placement predictions. In addition, we will examine whether the use of a proxy variable, which is foster care placement in our context, can potentially lead to racial disparity in child maltreatment predictions.

In particular, we find that the PRM tool is more accurate at predicting actual placement and child maltreatment for white children than for black children. Moreover, although out-of-home placement is a good proxy outcome for actual abuse risk, black children at high risk of placement are at much lower risk of severe abuse and neglect than similar white children. We estimate the true risk of placement for all children in the sample (N = 48,034) and find that black and white children have different underlying risk distributions. We then prove that the error rate balance criteria cannot be met due to the differences in actual risk distributions, *even when the PRM tool perfectly orders children according to their actual risk*.

This study's first objective is to examine how fairly the risk assessment tool predicts out-of-home placement for children across racial groups, particularly African-American vs white children. We begin by computing all the standard fairness metrics for each racial group and compare one against another. These are illustrated in *Figure 2*. The red bars represent 95% confidence intervals.

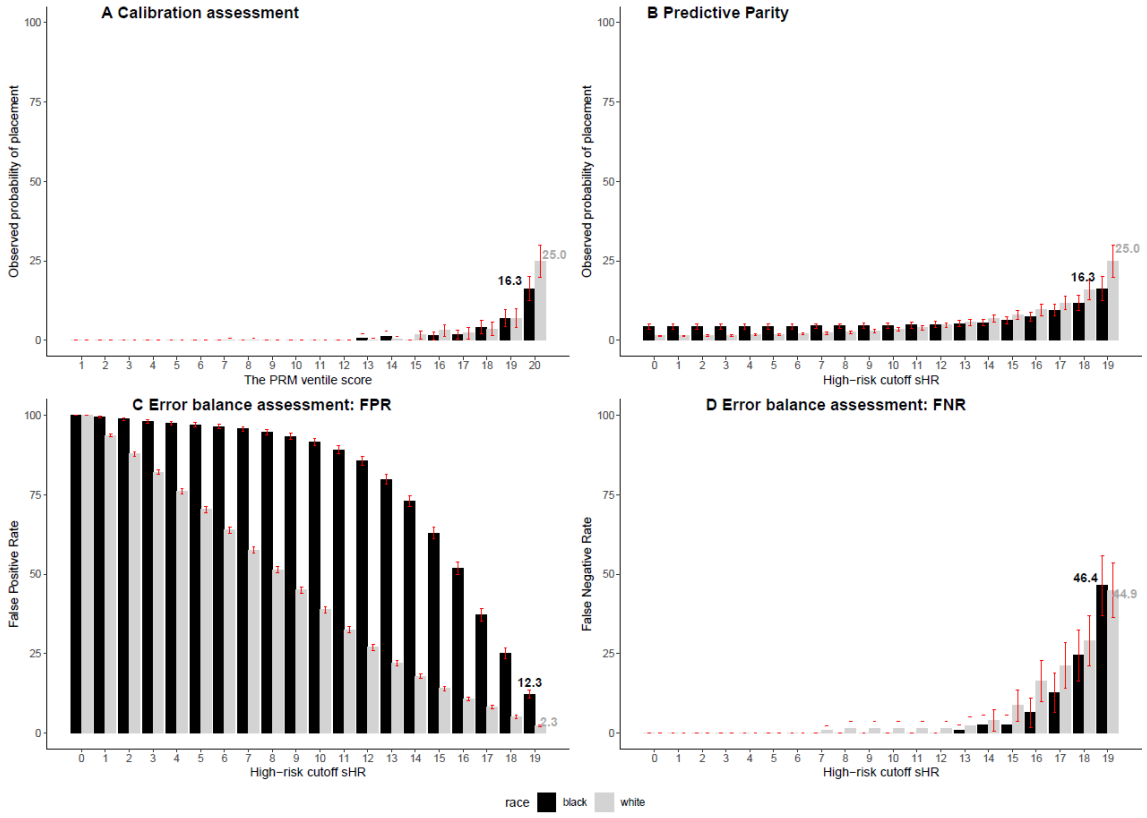

Figure 2: Empirical assessment of fairness criteria for the predictive risk tool used in child welfare. (A) Vertical axis represents the percentage of children who are placed in a foster care home by age three conditional on their predicted score. (B) Vertical axis represents PPV (positive predictive value) or the percentage of children who are placed in a foster care home by age three conditional on their predicted score exceeding the high-risk cut-off $s_{HR} \in \{0,..,19\}$ *measured on the horizontal axis. Everyone who scores above the cut-off is classified as high risk. Those at or below the cut-off are classified as low risk. (C) Vertical axis represents the percentage of children whose scores are higher than the chosen high-risk threshold* $s_{HR} \in \{0,..,19\}$ *conditional on not being placed in a foster care home by age three (False Positive Rate, FPR). (D) Vertical axis represents the percentage of children whose scores are not greater than the chosen high-risk threshold* $s_{HR} \in \{0,..,19\}$ *conditional on being placed in a foster care home by age three (False Negative Rate, FNR).*

By looking at *Fig. 2A*, we see that the tool is approximately equally well-calibrated at almost every score level except for the 5% highest risk (i.e., risk score 20). There are almost no children with scores 1 to 12, regardless of race, ending up in foster home care by age three. As the score increases, the observed frequency of placement increases for both black and white children, indicating that the tool is predictive of placement out-of-home. However, white children who score 20 are more likely to be placed out-of-home than black children predicted at a similar risk and the difference is statistically significant. In other words, the algorithm is better at predicting foster care placement for white than it is for black children. This is a critical observation since only the highest risk children

will be eligible to access the most intensive services. Although more than half of children who score 20 are black (Table 1), only 16% of them end up being removed from home by their third birthday. In contrast, white children account for approximately 43% (Table 1) of children who are predicted as at the highest risk, with 25% of them placed in foster home care in the next three years. Therefore, more black children are misclassified as at high risk of placement by age three compared to white children. Moreover, due to resource constraints, the service provider can only serve the top 5% highest risk children with the most intensive services to help the families cope with adversities and hopefully to reduce child maltreatment incidents before they occur. Naturally, based on the prima facie evidence in *Fig. 2*, we infer that the service providers might incorrectly allocate resources to less vulnerable black children, while missing out on opportunities to serve white children who are at higher risk of maltreatment. However, closer scrutiny is needed before we can accept this conclusion.

At a high-risk cut-off of 0, all children in the dataset are classified as high risk, so the observed probabilities of placement in *Fig. 2B* are population base rates. In other words, the bars for cut-off 0 in *Fig. 2B* tell us that if 5% of children (i.e., 2402 children) are randomly selected by the tool, there will be approximately 4% of the selected black children and 1% of the selected white children placed in foster care homes by age three. The likelihood of a black child being put in a foster care home is approximately three times higher than that of white children. On the other hand, if the random selection is among children who score 20, then 25% of white children and 16% of black children will be placed in care. Thus, compared to cut-off 0, choosing cut-off 19 shows a significant improvement in the model's predictive power and also reduces racial disparity. However, racial disparities in out-of-home placement outcomes still exist even at high-risk threshold 19.

Recall that the false-positive rate (FPR) refers to the proportion of non-placed children classified as high-risk, while the false-negative rate (FNR) is the proportion of placed children classified as low-risk. In the top 5% risk group, the likelihood of a non-placed black child being classified as high-risk is nearly five times that of white children (*Fig. 2C*). Meanwhile, a placed black or white child both have a similar likelihood of being assessed as low risk. Overall, the error rate balance criterion cannot be met for any cut-off (except 0), since false-positive rates are always unbalanced across races (*Fig. 2C*).

Another interesting observation from *Fig. 2* is that the conditions of predictive parity, equal false positive and equal false negative rates are not satisfied simultaneously at any cut-off. This is due to the placement prevalence rates being dissimilar across racial groups. Specifically, the rate of out-of-home care for black children is approximately 4%, whereas the white children placement rate is just above 1%. Recall that Chouldechova

(2017) asserts this incompatibility result, which was discussed in the previous section and proved in Appendix A. *Fig. 2* confirms that the PRM tool cannot satisfy all fairness criteria at the same time given the large discrepancy in foster care placement rates across racial subgroups. More specifically, the PP and FPR criteria are violated at a high-risk cut-off of 19 while the FNR condition is met. We are particularly interested in this cut-off since children in score bin 20 are eligible to receive the most intensive services from the service providers. Recall from the previous section that unbalanced FPR or FNR across race can lead to the racially disparate impact of tool use (Chouldechova, 2017). In the context of the COMPAS tool, which is used by judges to make bail decisions where offenders with a high risk of recidivism will receive heavier penalties than low-risk ones, Chouldechova (2017) uncovered that, on average, black defendants who have a higher base rate of recidivism, are more likely to receive a heavier penalty than white defendants, both among recidivists and non-recidivists. We learned from *Fig. 2* that FPR is statistically different in risk score bin 20, whereas FNR is fairly similar across race. Therefore, using the disparate impact analysis of Chouldechova (2017), non-placed black children on average receive more intensive services than non-placed white children. However, placed black children are not more likely to receive more intensive services than placed white children.

In the previous section we explained that the LASSO predicted probabilities are biased due to its functional form, which adds regularisation terms to the loss function. *Fig. 3* shows the empirical distribution of LASSO predicted probabilities of foster care placement across race. LASSO predicted probabilities are biased estimates of true risks. Black children's LASSO predicted probability distribution is approximately symmetric and unimodal, while that for white children is skewed right. Specifically, a greater proportion of white children are predicted as being low risk of future placement than black children. We notice that the range of LASSO estimated probabilities in *Fig 3* is from 0 to 1. The LASSO model predicts some children as having probabilities of placement as high as 90%, which seems unrealistic given that the average placement rate of the score 20 group is only about 21%. In other words, the model's level of certainty in predicting foster care placement seems excessively high. Since the LASSO algorithm is a regularised method, its predicted probabilities are mis-calibrated compared to regression-based methods such as logistic regression. This phenomenon is also referred to as prediction bias. One way to check this is by calculating the mean of the LASSO predicted probabilities for black and white children and compare those results against black placements base rate and white placements base rate. Black placements base rate is approximate 4%, whereas white placements base rate is 1%. However, the average of LASSO predicted probabilities for black is 0.48 (48%), while that for white children is 0.17 (17%). Thus, the mean of LASSO estimated risks of placement does not equal the

placement base rate, regardless of race. In other words, LASSO predicted probabilities exhibit significant non-zero prediction bias. However, the bias is introduced to (hopefully) reduce variance and improve out-of-sample predictive accuracy. Ideally the PRM tool should still rank individuals the same way as the true-risk ranking.

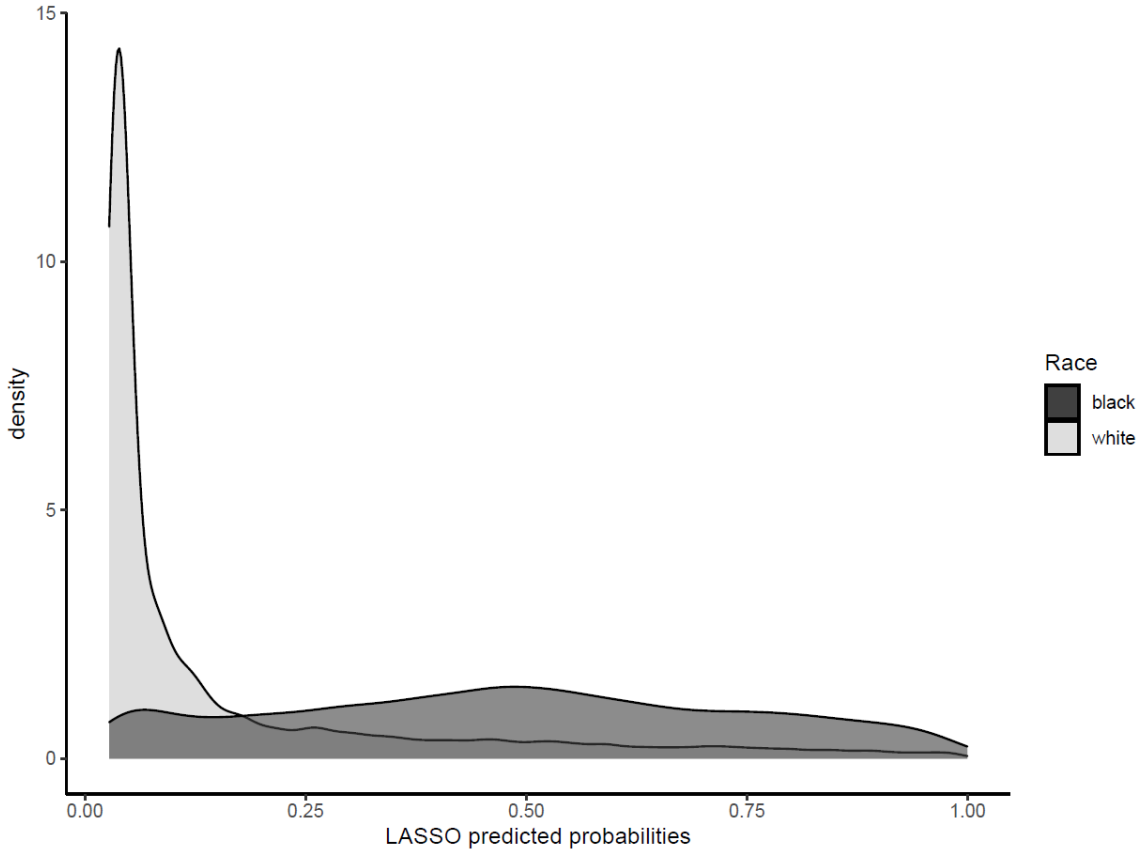

Figure 3: Empirical distribution of LASSO predicted probabilities produced by the studied PRM tool.

To estimate the true risk of foster care placement, we can regress the LASSO predicted probabilities against out-of-home placement outcomes using the sample of 48,034 children. Moreover, it would be interesting to know if the actual placement risks are different across races. We therefore include a dummy variable to indicate if the child is black and the interaction terms between the race dummy variable and LASSO predicted probabilities. Thus, the probability, *P*, of being placed in foster home care for a child with LASSO estimated probability P_{Lasso} , is defined as follows:

$$P = \Pr(Y = 1 | P_{Lasso}, B, B \times P_{Lasso}) \quad (1)$$

where Y is the out-of-home placement indicator variable (with Y=1 indicting that the child is placed in foster care by age three) and B is the Black dummy variable. To estimate the equation (1), we construct the logistic regression:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 P_{Lasso} + \beta_2 B + \beta_3 B * P_{Lasso} \quad (2)$$

$$\frac{P}{1-P} = e^{\beta_0+\beta_1 P_{Lasso}+\beta_2 B+\beta_3 B*P_{Lasso}} \quad (3)$$

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 P_{Lasso}+\beta_2 B+\beta_3 B*P_{Lasso})}} \quad (4)$$

Logit regression output

	Dependent variable:
	Placement by age three
Lasso probabilities	7.4***
	(0.2)
Black	0.1
	(0.3)
Black*Lasso probabilities	-0.7*
	(0.4)
Constant	-7.7***
	(0.2)
Observations	48,034
Log Likelihood	-2,946.9
Akaike Inf. Crit.	5,901.9
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Logistic regression on Black variable (i.e., if the child's race is black) and Lasso predicted probabilities and interaction terms between race dummy variable and LASSO predicted probabilities for the full sample (N = 48,034). Note: We include race in the conversion function from Lasso probabilities to actual risk as we want to see if the estimated true risks differ by race.

Table 2 implies that the same LASSO probability has different risk implications for different races. The race related coefficients on the interaction term in (4) are statistically significant which tells us that actual placement risk differs by race. Knowing this is the case, we separate the data set into the two race groups and run the logistic regression separately for each sub-sample.

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 P_{Lasso})}} \quad (5)$$

We estimated black children's true risk of placement using the function (5). A similar exercise is done for white children. The model results are in Table 3 in Appendix D. The estimated value for coefficient β_1 in (2) is positive and significantly different from 0 when regressing LASSO predicted probabilities against placement outcome for both black and white children subgroups (Table 3). Thus, LASSO predicted probabilities are predictive of out-of-home placement. However, black children's actual risk of placement distribution is dissimilar to that of white children. *Figure 4* shows LASSO predicted probabilities on X-axis and calibrated placement probabilities on Y-axis. We can see from *Fig. 4* that eligible black children are at significant lower risk of actual placement than similar white children.

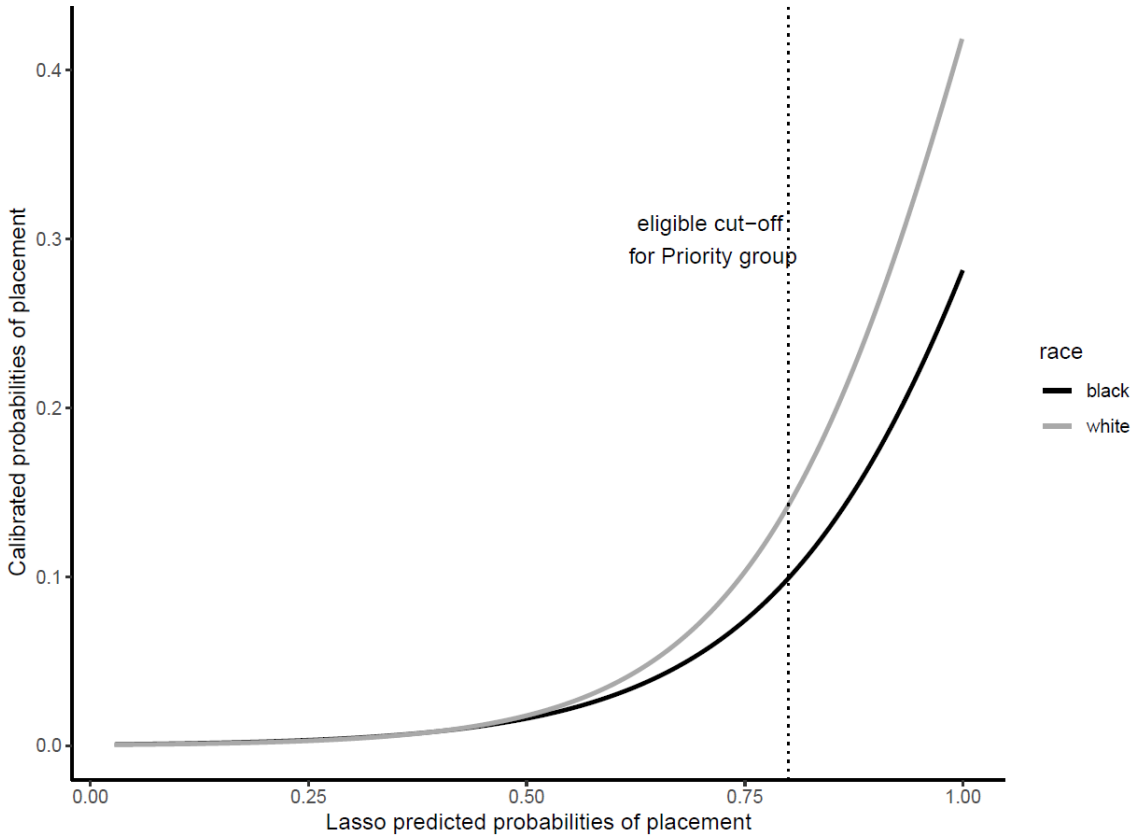

Figure 4: Estimated placement risk versus actual risk of placement, by race

Figure 5 shows the empirical distributions of true risk. In contrast to *Fig 3*, the distributions of true risk skew right for both racial groups. However, we can see that the underlying placement risk distributions continue to look different across race, though not as significantly as the LASSO risk distributions in *Fig 3*.

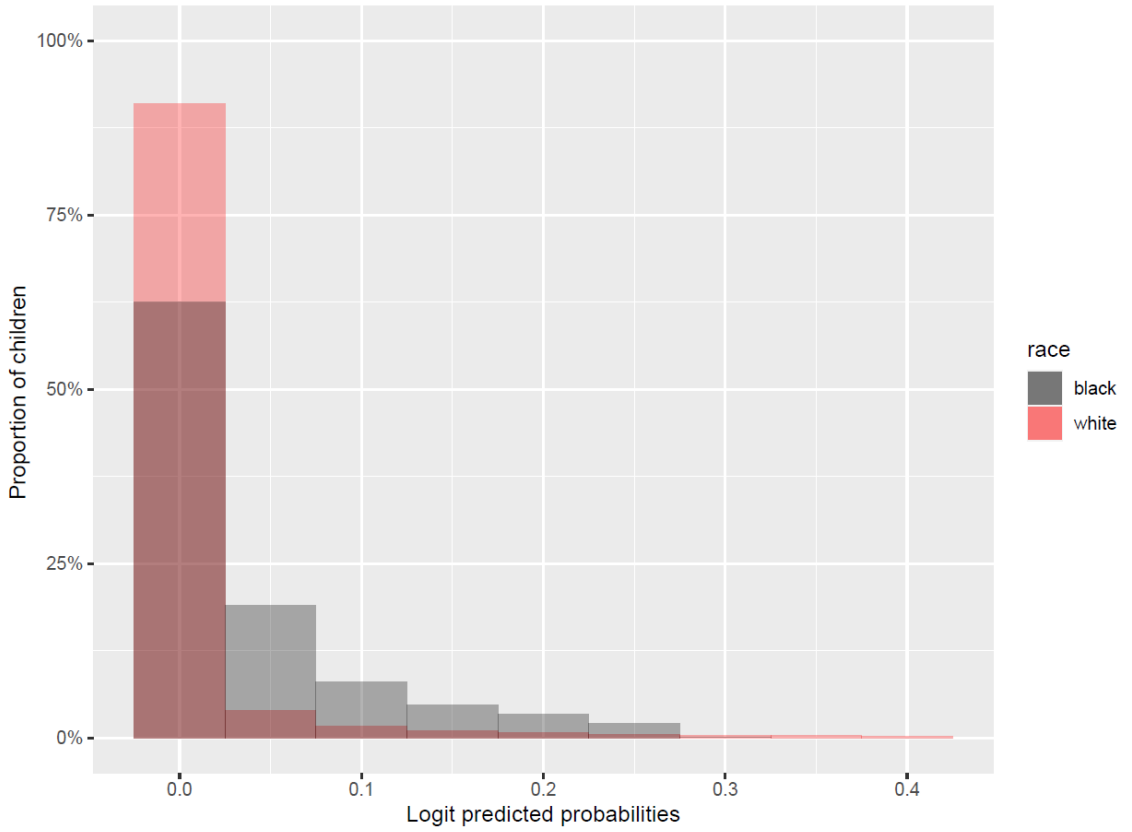

Figure 5: *Empirical distribution of estimated true risks by racial group (Histogram)*

As defined previously, to achieve unbiased prediction the average of predicted probabilities needs to be equal to the prevalence rate (i.e., the percentage of placed children in the data set). On average, 4% of all black children are placed in foster care home by their third birthday, whereas for white children the rate is 1%. The model (2) predicts on average that white children have 1% probability of being removed from home. Similarly, black children are predicted by model (2) to be removed at a similar rate to their population removal rate which is 4%. Thus, we get unbiased predicted probabilities of placement by constructing logistic regression (2) for each race group. *Fig. 5* shows that no white children have more than a 45% chance of being placed in foster care by age three, whereas for black children, the maximum risk is less than 30%.

Table 4 in Appendix E provides the implied score cut-offs for actual placement risks. Children in the top 5% highest risk of placement now have their estimated logistic probabilities greater than 0.122 (i.e., α = 0.122) (Table 4).

These distributions of actual placement risk for black and white children contain useful information to help us understand why, and how, ERB fails in *Fig. 2C*. As shown in

Appendix B, we can visualise the theoretical ERB metrics[4] using the estimated probability distribution functions for the actual risk of placement (*Fig. 7*). The cumulative distribution function (CDF) curves are constructed using the probability density function of true risk in *Figure 6* in Appendix H. They are displayed in *Figure 7*. The FNR is calculated as B/(A+B), where A is the area above the CDF curve between horizontal lines at heights $F_g(\alpha)$ and 1 and B is the area above the CDF curve and under the horizontal line at height $F_g(\alpha)$ (where F_g is the CDF for group g). The FPR is equal to C/(C+D) where C is the area below the CDF between horizontal lines at heights $F_g(\alpha)$ and 1, while D is the area below the CDF up to the horizontal line at height $F_g(\alpha)$. Proof of these claims can be found in Appendix B. For each race group's CDF in *Fig. 7* the areas A, B, C and D for α=0.122 are shown.

Using the results from Appendix B, we can now calculate FPR and FNR for each racial group.

We calculated that $A_w \approx 0.007$ & $A_b \approx 0.021$, $B_w \approx 0.006$ & $B_b \approx 0.019$.

And $C_w \approx 0.023$ & $C_b \approx 0.089$, $D_w \approx 0.964$ & $D_b \approx 0.871$.

Thus

$FPR_w \approx 0.023$ & $FPR_b \approx 0.093$ and $FNR_w \approx 0.462$ & $FNR_b \approx 0.475$

Thus $FNR_w < FNR_b$ & $FPR_w < FPR_b$ at the top 5% highest risk. False negative rates are not significantly different across race, while the false positive rate calculated for black children is approximately 4 times higher than that for white children. ERB values derived from the actual risk distribution function (i.e., CDF curves of true placement risk) are consistent (but not identical) with *Fig. 2C* and *Fig. 2D* which were calculated using LASSO predicted risk scores and the realised outcome of the sample (see Footnote 4). However, using the estimated true risk distributions help us understand why the FPR metrics differ as they do across races. If the predicted (true) risk distributions were identical for black and white children, the PRM tool would satisfy the error rate balance conditions at any given choice of high-risk threshold. Conversely, whenever the distributions vary by race, at least one of the ERB conditions will likely fail even if the PRM tool correctly ranks children by risk.

[4] These "theoretical" ERB metrics are not based on the actual (realised) outcomes of this particular sample of children. They measure the expected FPR and FNR for a random sample of people drawn from the estimated risk distribution. Hence, our calculations below do not exactly match the FPR and FNR in *Fig. 2*.

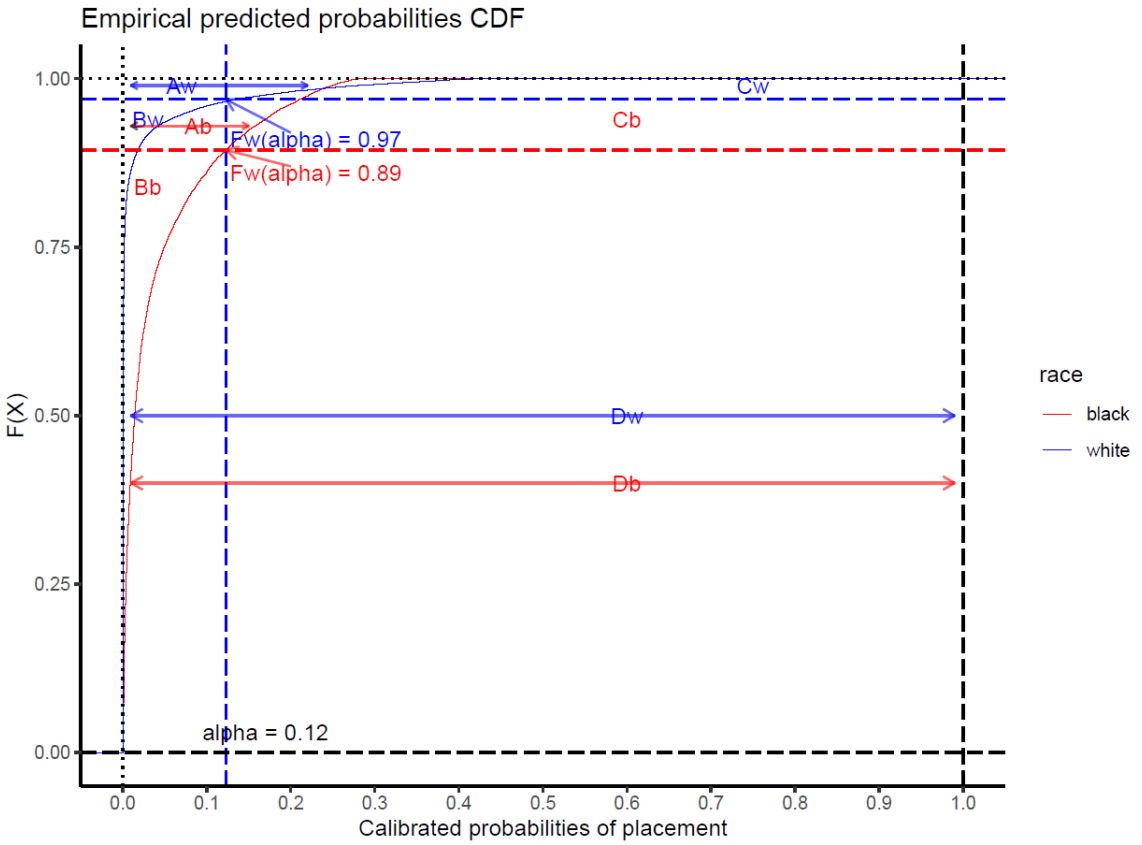

Figure 7: Empirical cumulative distribution function by racial groups

To better understand why the shapes of the actual risk distribution can affect the satisfactory ERB condition in fairness assessment, we employ some artificial CDF curves to demonstrate. *Figure 8A* shows a hypothetical situation in which black children's true risk distribution is uniform while white children's one is right-skewed. *Figure 8B* shows another hypothetical scenario with black children's underlying risk distribution being unimodal whereas the white children's one skews right. In each hypothetical scenario, the base rate is the same across races – the area above the red CDF (up to height 1) is the same as the area above the blue CDF. Obviously, the FPR and FNR calculated for black and white children using the shapes of CDFs as described in *Fig. 8* will be different across race. More specifically, in these two scenarios, black children will have larger FPR and FNR. In other words, if the underlying risk distributions are relatively dissimilar across race, the ERB condition will fail since the FPR and FNR conditions cannot both be met.

The key point drawn from the CDF analysis as discussed above is that ERB metrics will differ if underlying risk distributions differ, even if these risks are perfectly estimated. In other words, ERB criteria have almost nothing to do with how well (or badly) the PRM predicts risk. Moreover, the shape of the CDF curves directly affects the value of FPR

and FNR. For instance, for children with a specific placement base rate, when the CDF curves represent left-skewed distribution of placement risk (i.e., the blue curves in *Fig. 8*), FPR and FNR will be lower compared to when the risk distribution is unimodal or right-skewed (i.e., the red curves in *Fig. 8*). In other words, when the CDF curve moves further away from the top left corner in *Fig. 5*, FPR and FNR increase. This observation helps explain why $FNR_w < FNR_b \; and \; FPR_w < FPR_b$ as in *Fig. 2*. It is because the shape of true risk distribution for black children lies further away from the top left corner (*Fig. 5*) compared to that for white children.

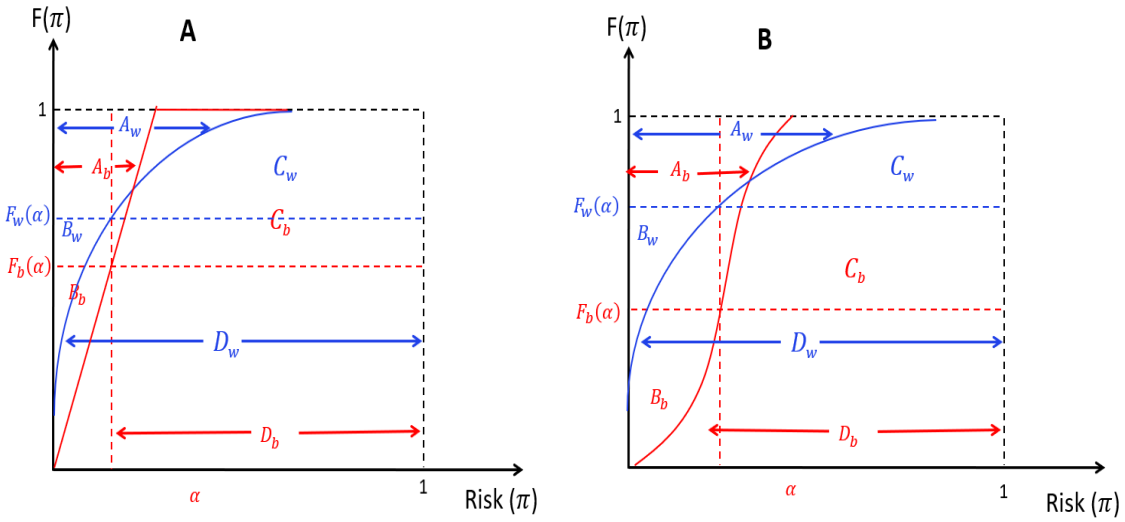

Figure 8: Hypothetical risk distribution functions for black and white children. A) shows right-skewed risk distribution for white and uniform risk distribution for black children. B) shows similar risk distribution for white as picture A but unimodal risk distribution for black children.

There is a close relationship between ERB and AUC since the ROC curve is a graphical depiction of the TPR (or 1-FNR) and FPR combinations for different cut-offs (α), so it captures the same information as the ERB data in different form. By varying alpha, we compute TPR and TNR using the area formula in Appendix B and plot the (FPR, TPR) combinations for each racial subgroup (*Fig. 9*). In particular, the estimated AUC for white children is almost 95% (95% CI: 0.9345-0.9647), whereas that for black children is just under 83% (95% CI: 0.7964-0.8661). This difference in AUC for black and white children is statistically significant at a 95% confidence level.

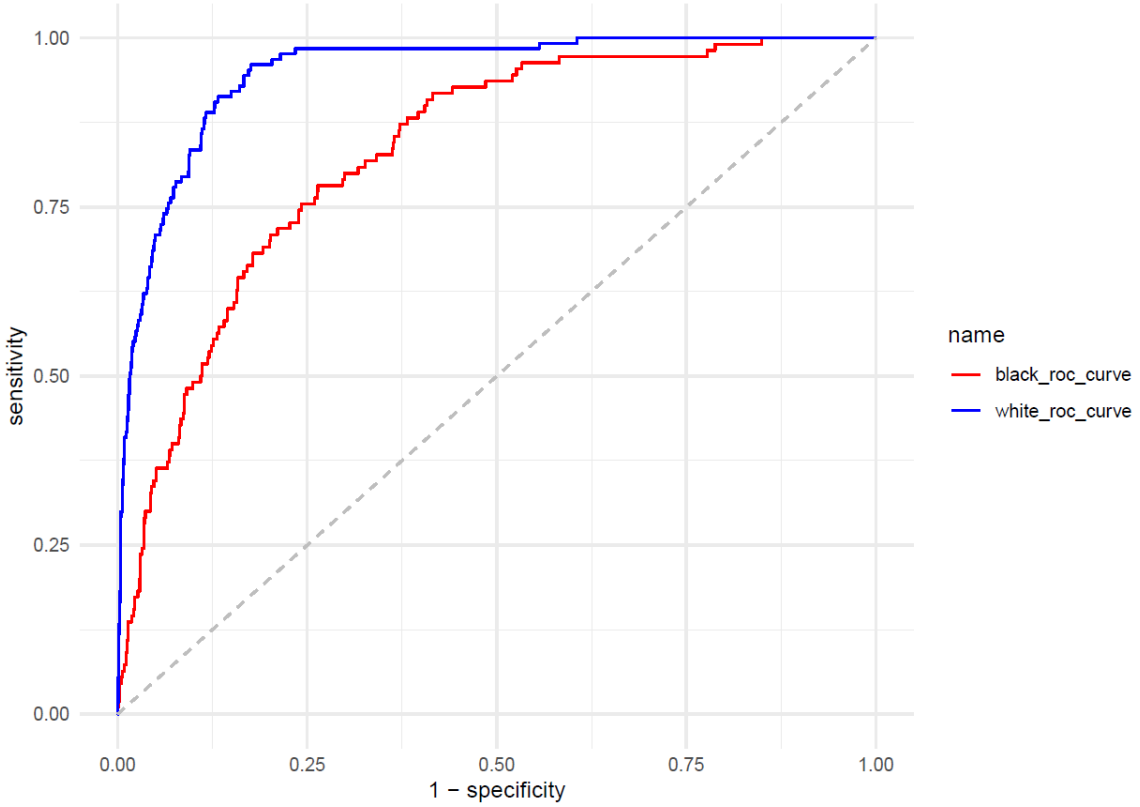

Figure 9: Placement ROC curves for black and white children (test set)

Recall from the previous section that ROC curves show how discriminating estimated risks are. In other words, it shows the tool's ability in differentiating non-placed children from placed children based on their placement risk. Since the ROC curves basically summarise the ERB numbers in another form, we can expect a different AUC for black and white children given their different risk distributions, and this is confirmed in *Figure 9*.

In addition, our analysis in Appendix B shows that the ERB basically reflects differences in the underlying placement risk distributions of the two race groups. However, it tells us very little about how well the tool predicts true risk of placement. When we see different AUC's for black and white children, we don't know if this is just due to (i) different distributions of true risk, and/or (ii) the risk estimates for black children being more "noisy". Even if the tool predicted risk perfectly (i.e., perfectly ranked everyone in terms of underlying risk), there would still be a huge difference in AUC and hence violation of ERB due to the dissimilarity in the underlying placement risk distribution for black and white children. Recall that algorithmic bias is about assessing whether the tool is better at assessing risk for one group than another. In this sense, AUC or ERB plays little role in evaluating how fair the tool is. Although ERB constantly appears in the literature as a

means to assess algorithmic fairness, our analysis shows that it is mis-used. As mentioned above, ERB is more relevant when assessing the disparate impact that the decision causes.

In order to assess the fairness of the PRM tool in predicting placement risk, it is crucial to understand how accurate the tool is at ranking children according to their true risk. Recall *Fig. 4*: this shows systematic differences (between black and white) in the model's ability to predict risk. Specifically, the PRM tool overstates black children's true placement risk relative to white children's true risk. Among high-risk children whose LASSO predicted probabilities are greater than 0.8 (Table 4), white children are at significantly higher risk of placement than black children. Table 2 shows that the differences between black and white children's true risk of placement are statistically significant.

When we see significant divergence in calibration (between black and white children) at score 20 (*Fig. 2A*), this could be due to (i) different conditional risk distributions over the score 20 probability range (e.g., most white children at the top end and most black children at the bottom) and/or (ii) systematic differences (by race) in the model's error in estimating risk. Black and white children who score 20 have a similar conditional risk distribution, as shown in *Fig. 10*. It follows that the placement rate difference at score 20 is (mostly) due to the different mappings from LASSO probabilities to true risks. In particular black children who are predicted as high-risk of placement by the algorithm indeed have relatively lower risk of placement compared to similar white children. Put differently, black children's true risk of placement is not correctly predicted compared to white children's. Thus, the PRM tool is more accurate in predicting actual placement risk for white children than it does for black. Besides, we learn from *Fig. 4* that the tool mis-ranks the children's true risk. Specifically, a black child whose LASSO predicted probabilities is just above the high-risk cut-off of 0.8 are at significantly lower risk of placement than a white child whose estimated risk is just under the high-risk threshold.

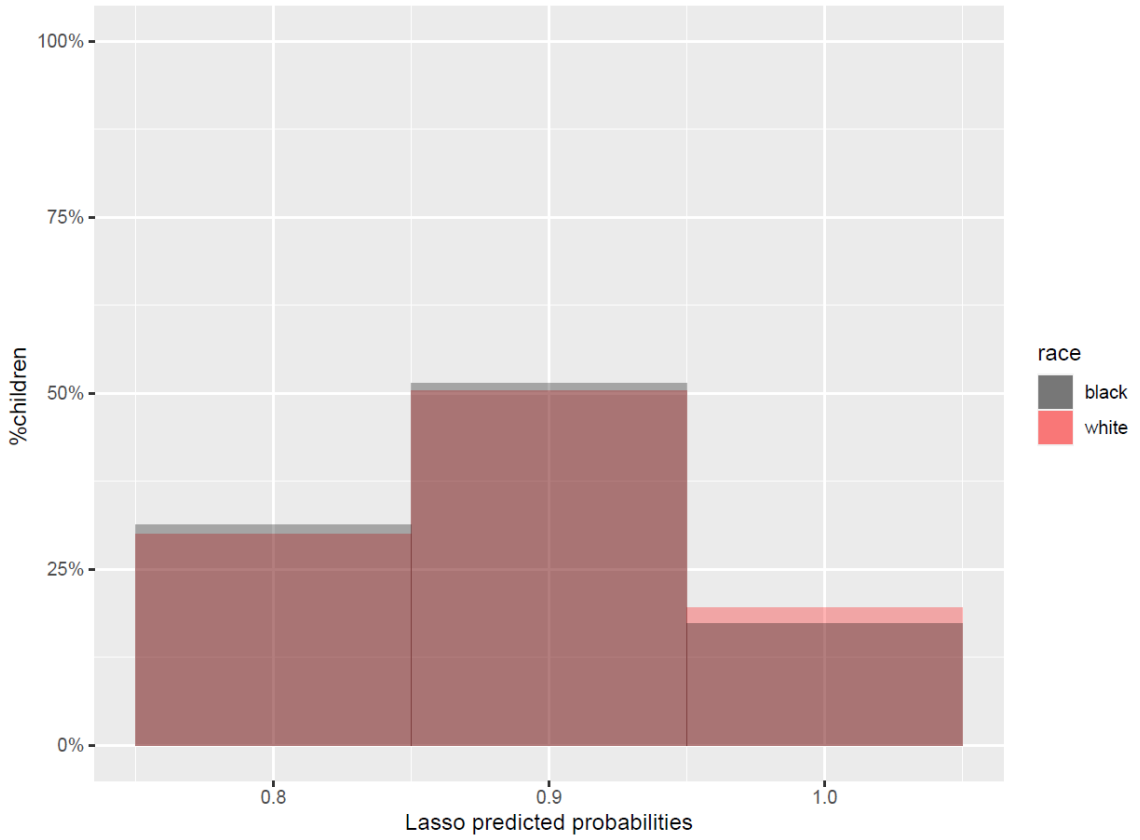

Figure 10: Conditional estimated risk distribution for black and white children who score 20

So far, we have learned that the PRM tool does not always rank children according to their true risk of foster care placement. It is also important to know how the algorithm orders children relative to their true risk of abuse and neglect. We use sensitive death, which is defined in the previous section as being due to accident, violence or maltreatment, as a proxy for severe abuse and neglect. Replicating the regressions that we did for out-of-home placement, Table 5 in Appendix F shows that sensitive death risk also differs by race, and the differences are statistically significant. We then estimate black children's true risk of abuse using the function (5) but with the dependent variable now being the probability of the sensitive death outcome. A similar exercise is done for white children. The model results are in Table 6 in Appendix G and the estimated functions are depicted in *Figure 11*.

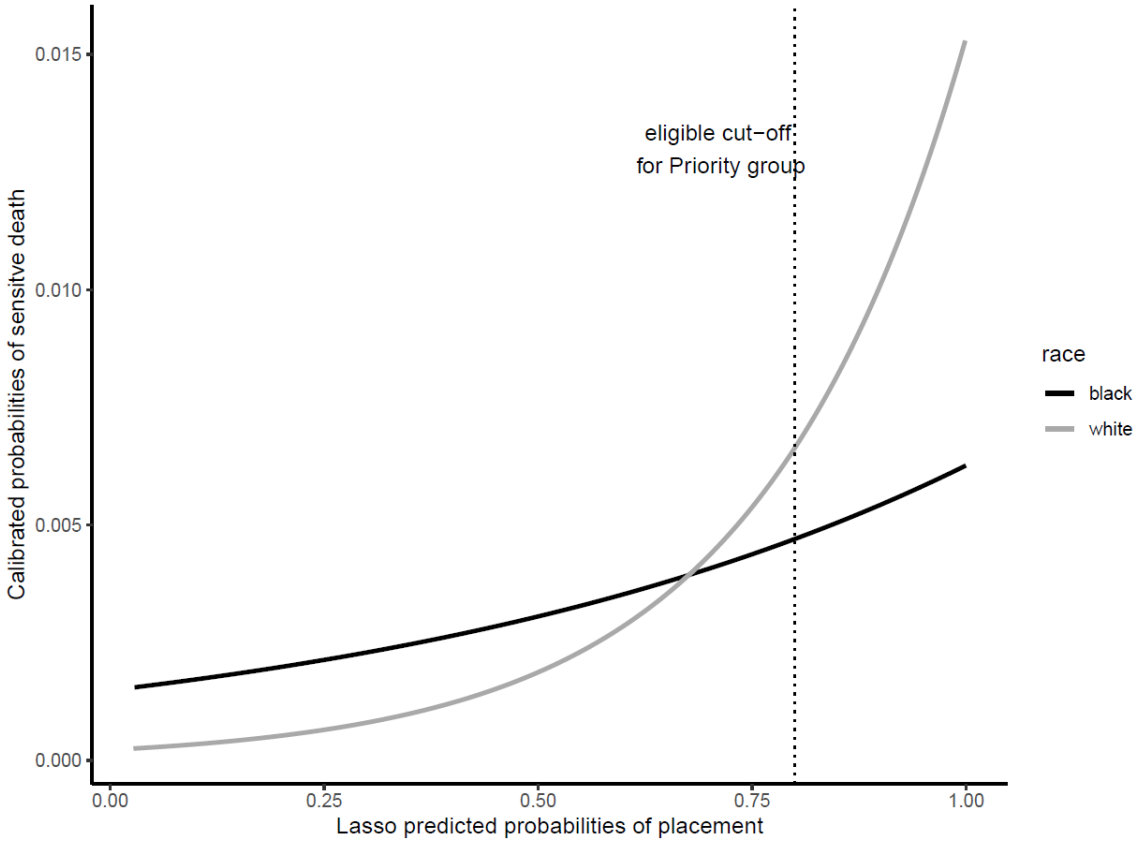

Figure 11: Estimated placement risk versus actual risk of sensitive death, by race

The PRM tool can predict true abuse risk reasonably well since both curves have positive and significant slopes *(Fig. 11*). However, black children who are estimated by the PRM tool as having low-risk of placement have a higher risk of sensitive death compared to similar white children. Conversely, the opposite bias is found for children whom the algorithm predicts as at high-risk of foster care placement. The grey line is a steeper slope than the black line (*Fig. 11*) which shows that compared to low-risk of placement white children, high-risk white children have significant higher risk of abuse. Meanwhile, the difference in terms of abuse risk, of low-risk and of high-risk black children are less dramatic. Therefore, the model's ability to discriminate levels of abuse risk for white children is better than it is for black children.

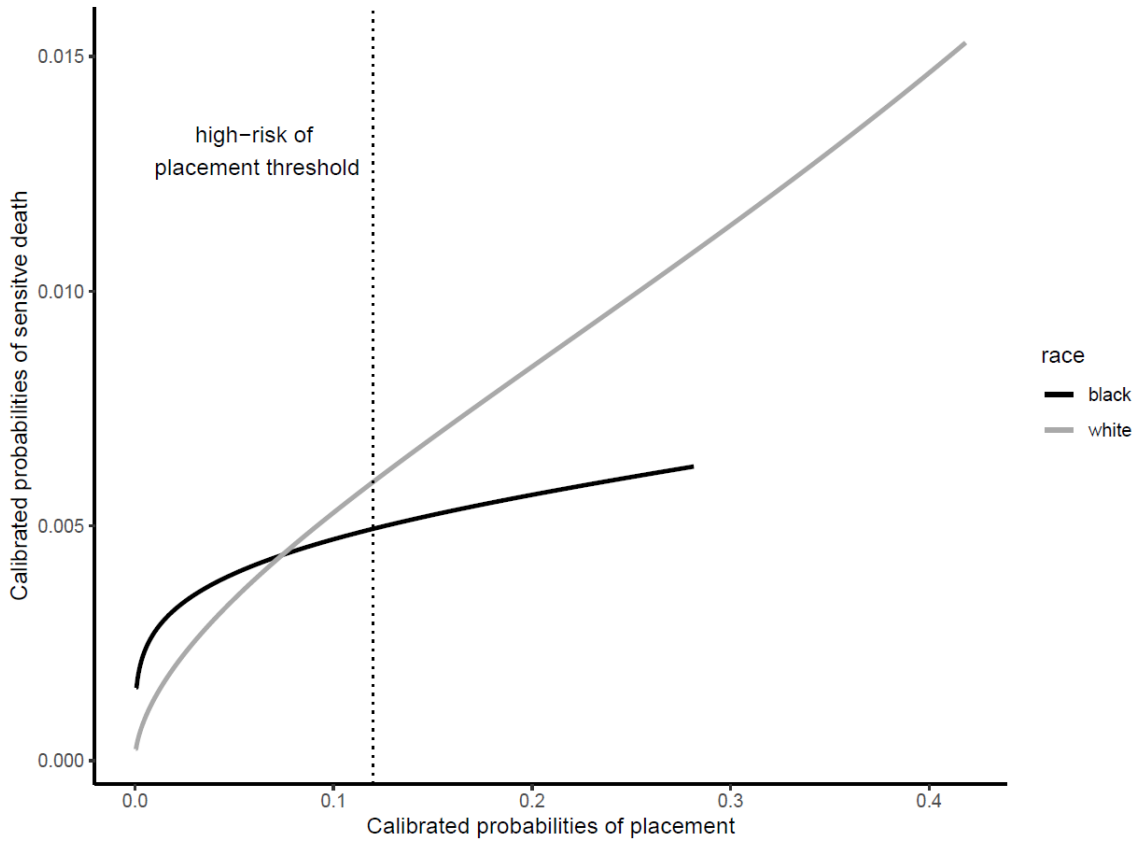

Figure 12: Actual placement risk versus actual abuse risk, by race.

In sum, an important lesson drawn from our analysis is that ERB or AUC plays a relatively small role in assessing the fairness of the PRM tool. Because even if the algorithm is perfectly precise at predicting placement, due to the difference in underlying actual placement risk distribution, the ERB or AUC criteria will not be satisfied. Therefore, to better assess algorithmic fairness, we need to understand if the tool mis-ranks people according to their true risk. In the context of the studied PRM tool, we find that it is not equally accurate in predicting placement and sensitive death risk for black and white children. In addition, the algorithm does not correctly rank children by their true risk of placement and abuse. More specifically, there is mis-ranking in the tool's predictions of children's true risk, such that black children eligible for the Priority service tier are at lower risk of placement and abuse than white children who are prioritised for the same service. This finding leads us to examine more closely the quality of placement as a proxy for child maltreatment. *Figure 12* shows the correlations between actual placement and abuse risk by race. Interestingly, the actual placement risk is a good proxy for abuse risk (both curves are upward sloping) but the strength of the correlation varies by race and level of risk (*Fig. 12*). At lower levels of placement risk, black children seem to have a higher chance of dying because of abuse and neglect. However, when the likelihood of placement increases, black children are placed not solely because they are at risk of

abuse and neglect, but perhaps for some other reasons. *Fig. 12* suggests that black children can be placed in foster care home even when their child maltreatment risk is the same as similar white children. Meanwhile, white children have to be at significant risk of abuse for them to be removed from home.

These results suggest that there is algorithmic bias both in the ranking of children by placement risk and also due to the proxy choice. At the high-risk cut-off of 0.8 (i.e., the current eligible threshold to receive the most intensive services), white children's estimated placement risk is approximately 0.14, while for black children that risk is relatively lower - approximately 0.1 (*Fig. 4*). Using *Fig. 12*, we determine the estimated sensitive death risk for black children whose placement risk is 0.1 is much lower compared to white with placement risk of 0.14. Therefore, for the eligible group, the two forces (i.e., mis-ranking in placement risk estimation and proxy bias) are compounding each other. Moreover, conditional on placement risk, black children suffer different risks of abuse and neglect compared to white children (*Fig. 12*). As a result, precise estimation of out-of-home placement risk still means being racially biased on child maltreatment prediction. The next section will discuss possible reasons for these disparities.

7. Mechanism of bias

According to findings from the previous section, the PRM tool does not correctly score children according to their true risk of out-of-home placement or child abuse. In particular, black children at lower risk of placement are more likely to be predicted as high-risk than similar white children. More importantly, among the high-risk of placement group as estimated by the tool, black children are at significantly lower risk of actual child maltreatment than white children. *Fig. 12* shows us that black children's placement outcome does not guarantee that they are genuinely at risk of child maltreatment. For white children, placement outcome is a better proxy than it is for black children. This section will provide possible answers to the two following questions: How do the disparities in placement risk occur for black and white children with the same risk of child maltreatment? Specifically, why are placed black children at a lower risk of abuse and neglect than placed white children?

Ultimately, the tool's objective is to precisely capture which children are at significantly higher risk of abuse and neglect due to their adverse circumstances. The service provider can intermediately support them and their families to avoid critical events from happening. However, measuring actual abuse and neglect is neither easy nor sometimes even possible, since the child welfare system does not have enough resources to triage all abused children. Even if they make enough effort to investigate all possible alleged abuse, it does not guarantee that the investigation outcomes are 100% accurate. Also, some abuse is unreported; thus, we might have missing data for some severe maltreatment. Abuse and neglect are very subjective unless the signature is apparent and universally accepted. Often, child maltreatment signs are subtle and can be easily misinterpreted. Due to the difficulty in measuring actual abuse and neglect, the tool's designers used out-of-home placement as a proxy outcome. It can be measured accurately and is correlated with current maltreatment as well as possible risk of further abuse and neglect requiring CPS involvement. Thus, a key assumption is made that children at high risk of foster care placement are also the children at significantly higher risk of abuse and neglect. In other words, the tool’s developers predict out-of-home placement while hoping that their tool is also good at predicting child maltreatment.

On the one hand, it seems reasonable to expect a positive and robust correlation between out-of-home placement and actual maltreatment, for several reasons. Firstly, investigators must substantiate the abuse allegation before deciding whether to remove the child. Moreover, the removal decision has also to be based on evidence of possible further risk of maltreatment. In other words, collective evidence of current abuse and

neglect and the future threat of maltreatment has to be collected before CPS workers can decide to remove the child from the home.

On the other hand, it has been suggested broadly in the literature that placement outcomes have differential abuse risk implications for black and white children. There are several possible channels for such disparities to arise. First, out-of-home placement might be subject to racial bias (Chibnall et al., 2003; Font et al., 2012). A recent study suggests that paediatricians are more likely to evaluate and report black children with fractures as suspected maltreatment than similar white children, especially among toddlers (Lane, 2002).). In this study, Lane and his colleagues also indicate that race is often employed to diagnose child maltreatment in the hospital context further. Physicians frequently order skeletal check-ups for black children to ensure that they do not miss any signs of abuse. However, the same action is considered unnecessary for white children. Therefore, the authors suggest that physical abuse among white children is easily neglected and often under-reported (Lane et al., 2002). Several other studies also support the hypothesis that a suspected case of physical abuse or sexual abuse involving black children is more likely to be referred to the CPS system than a similar case involving white children (Ards et al., 2003; Bartholet, 2009; King et al., 2017; Lavergne et al., 2008). These results suggest that black children with maltreatment allegations are more likely to be referred to the CPS system than their white counterparts. Another study of drug use during pregnancy found that healthcare staff were more likely to refer black mothers to CPS than white mothers, even though they were equally likely to have a positive drug test (Chasnoff et al., 1990). Because reporting is a gateway to the CPS system, if an abused white child is never referred to the system before they face a critical incident, their abuse might never be found until they are severely injured or even die. However, a non-abused black child who is constantly reported might make people think they are at risk of maltreatment; thus, they might be placed in foster care homes because of racial bias, not because of their actual risks.

According to a recent systematic review regarding the over-representation of black children in the child welfare system, African-American children are more likely than white children to be involved at multiple stages of contact with child protective services (Cénat et al., 2021). Several studies found that maltreatment allegations involving black youth are more likely to result in the case being screened-in and investigated than maltreatment allegations involving white youth (Fluke et al., 2003; Font et al., 2012; Harris & Hackett, 2008). Due to racial bias and discrimination, black children are more likely to be put in foster home care and have a more extended stay than white children who have a similar risk of abuse and neglect (Cénat et al., 2021).

Some studies have found that race does not directly affect the risk assessment process; however, when it comes to decision-making regarding abuse cases, race emerges as a determining factor (Dalgleish, 2006; Dettlaff et al., 2011; Rivaux et al., 2008). In particular, Dagleish (2006) employs a "signal detection framework" to show that decision-makers use information associated with the present case to evaluate risk. However, the author suggests that decision-makers' own experiences, such as perception of race, do influence their final action (Dagleish, 2006). More importantly, their findings suggest that case workers apply different thresholds to substantiate decisions to remove black and white children. More specifically, case workers set a higher risk threshold to remove white children. Thus, white children removed from their homes are at significantly higher risk of abuse and neglect than similar placed black children (Dagleish, 2006; Rivaux et al., 2008). Put differently, African American children in foster care homes are at lower risk of child maltreatment than white children because the case workers apply a lower decision threshold to remove them. As a result, maltreated white children may be under-represented, whereas non-abused black children are likely over-represented in foster care homes (Detlaff et al., 2011).

Second, a lack of cultural competence among child welfare workers and the racial differences between families and their case workers could also contribute to the over-representation of low-risk of abuse black children in the CPS system. In one study, the authors suggest that white case workers did not have enough exposure to minority cultures and norms and a good understanding of disciplinary practices (Chasnoff et al., 1990). Moreover, reporters often base their referral decision on their pre-existing beliefs of what constitutes harm and maltreatment (Font et al., 2012). Thus, acceptable physical punishment in African-American families might look like abusive behaviours in the eyes of reporters who come from different backgrounds or cultural beliefs. Perhaps, it explains why black children are seen as more vulnerable than white children.

Finally, misperception toward black parents could contribute to the over-representation of black children in foster care homes while their risks are not significantly high. Although child protective services are set up to protect children and empower all families alike to do better, out-of-home placement sometimes seems to work as a punitive tool to punish minority parents for their failures in raising their children. A few studies found that child welfare workers seem to show sympathy towards white parents and believe that white parents can change and adopt non-abusive educational practices (Chibnall et al., 2003; Clarke, 2011). On the other hand, McCallum and Cheng (2016) show that case workers have less trust in Black parents' ability to change and work on new non-violent parenting approaches. show that case workers have less trust in black parents' ability to change and work on new non-violent parenting approaches. Moreover, black parents are

assumed to be irresponsible, uneducated and difficult to rehabilitate. As a result, white children are more likely to receive in-home services after being substantiated for child maltreatment than similar black children (Needell et al., 2002; Lavergne et al., 2008).

In sum, child removal decisions could be biased due to the collective effect of all the factors explained previously. In particular it is more likely to be upward biased for black children and downward biased for white children. Black children with injuries and neglect-like situations are more likely to come to the attention of both reporters and the CPS system, whereas similar white children might be overlooked. Moreover, case workers often set a higher threshold to remove white children from home (Dettlaff et al., 2011; Rivaux et al., 2008). This implies that white children are only taken away from their parents when there are apparent signs of severe abuse and neglect. Conversely, black children's risks are frequently overestimated, and they are more likely to be taken away from their parents for less severe abuse. Notably, even at the same level of abuse and neglect, white children are more likely to receive in-home services to preserve family integrity. Thus, out-of-home placement risk is a better proxy for child maltreatment for white than for black children.

8. Conclusion and Discussion

In closing, the PRM tool of interest is not equally accurate in predicting out-of-home placement for black and white children. Racial bias exists in the PRM tool's placement predictions studied in this paper. In particular we argue that there is a mis-ranking in the tool's predictions of children's true risk, such that some black children eligible for the Priority tier service are at lower risk of placement and abuse than some white children who are prioritised for the same service. In addition, we suggest that the proxy choice issue could also cause the algorithmic bias we detect. That is, conditional on placement risk, black children suffer a different risk of abuse and neglect compared to white children,ie., the two forces (mis-ranking in placement risk estimation and proxy bias) reinforce each other. The compound effect is evident amongst the eligible children. Even though placement risk is a good proxy for abuse risk for black and white children, the strength of the correlation varies by race and level of risk. More specifically, out-of-home placement is a better proxy for abuse for white than for black children. Furthermore, we prove that ERB or AUC plays a relatively small role in assessing the fairness of the PRM tool. Even if the algorithm is perfectly accurate at estimating placement risk, due to differences in underlying actual placement risk distributions, ERB or equal AUC criteria will not be satisfied.

Understanding that the choice of proxy variable can potentially lead to racial disparities in the model outcome predictions prompts us to seek a less biased proxy. Although the proxy bias issue can be fixed (Obermeyer et al., 2019), generating new proxy variables are relatively challenging and resource-intensive. It requires deep knowledge of the child welfare system, the ability to collect and retrieve the relevant data, and iterative experiments. The literature posits that white children are more likely to receive in-home services than black children even at the same level of abuse and neglect (Lu et al., 2004; Miller, 2008). We therefore suggest using a new target variable to indicate whether the child should receive either in-home services or out-of-home removal, after their allegations have been substantiated. This way, we can capture more children at high-risk of abuse and neglect. Due to limited data access, we did not test this new proxy variable to see if it reduces racial disparities in child maltreatment predictions produced by the PRM tool. However, this could be a subject of our future research.

Several limitations exist in our research, mainly due to our limited access to the relevant data. The implications and conclusions of the current research rely heavily on the out-of-home placement and sensitive death labels provided to us. If those outcomes are incorrectly labelled, our results might be affected. We were not able to verify the accuracy of the labelling task due to limited data access. Furthermore, we cannot retrain the model

using sensitive death to extract predicted probabilities and compare the ranking of children across models to see if children who are high-risk in the placement model are also high-risk in the sensitive death model. It would be interesting to see if children's rankings will be changed from the out-of-home placement model to the sensitive death one, because this comparison will help us understand further whether the current PRM tool which predicts home removals, is also accurate in identifying children who are at significant risk of severe maltreatment that could lead to mortality. Therefore, this is another avenue for potential future research.

APPENDIX A

In this appendix we prove that when base rates are different across groups, the recidivism prediction instrument (RPI) cannot achieve all three fairness conditions simultaneously. Specifically, we show that the following equation must hold:

$$\text{FPR} = \frac{p}{1-p}\frac{1-PPV}{PPV}(1-FNR) \tag{6}$$

Suppose:

N is total number of people

TP: the number of people with high scores ($S>s_{HR}$) who recidivate in the future (Y=1): the true positives.

FP: the number of people with high scores who do not recidivate in the future (Y = 0): the false positives.

TN: the number of people with low scores ($S\leq s_{HR}$) and Y=0: the true negatives.

FN: the number of people with low scores and Y=1: the false negatives.

Proof:

The base rate of recidivism is calculated as follows: $p = P(Y = 1|R = r) = \frac{FN+TP}{FN+TP+TN+FP}$

,

while $1 - p = P(Y = 0|R = r) = \frac{TN+FP}{FN+TP+TN+FP}$ is the prevalence of nonrecidivism.

Total number of defendants is: $N = FN + TP + TN + FP$

With false-negative rate is calculated: $FNR = \frac{FN}{FN+TP} = \frac{FN}{p\times N}$

And $PPV = \frac{TP}{FP+TP}$ & $FPR = \frac{FP}{FP+TN} = \frac{FP}{(1-p)\times N}$

Therefore,

$$\frac{1-PPV}{PPV} = \frac{1}{PPV} - 1 = \frac{TP+FP}{TP} - 1 = \frac{FP}{TP} \tag{8}$$

And the ratio between recidivism rate and nonrecidivism rate is:

$$\frac{p}{1-p} = \frac{TP+FN}{TN+FP} \quad (9)$$

And

$$1 - FNR = 1 - \frac{FN}{FN+TP} = \frac{TP}{FN+TP} \quad (10)$$

From (8), (9) and (10, we can derive (6):

$$\frac{p}{1-p}\frac{1-PPV}{PPV}(1-FNR) = \frac{TP+FN}{TN+FP} \times \frac{FP}{TP} \times \frac{TP}{FN+TP} = \frac{FP}{FP+TN} = FPR$$

$$\begin{cases} FPR_b = \frac{p_b}{1-p_b}\frac{1-PPV_b}{PPV_b}(1-FNR_b)(for\ Black\ people) \\ FPR_w = \frac{p_w}{1-p_w}\frac{1-PPV_w}{PPV_w}(1-FNR_w)(for\ White\ people) \end{cases} (I)$$

From the system of equation (I), one can see that the base rate plays an essential role in keeping the error rate balance across racial groups when the RPI satisfies predictive parity and equal FNR. However, it is also an uncontrollable factor to which most predictive models might be subject. Although the RPI meet the predictive parity criterion, there will be large discrepancies in error rate when the base rates are significantly different.

Appendix B

CDF proof:

Suppose we have two groups: White (w) and Black (b). There are $\boldsymbol{N_g}$ people in group g $\in \{\boldsymbol{w}, \boldsymbol{b}\}$**.** Let N = $\boldsymbol{N_w} + \boldsymbol{N_b}$.

A child is assigned outcome Y = 1, meaning having placement by age three, with probability π and Y =0 (e.g., not having placement by age three) with probability 1-π. For people in group g $\in \{\boldsymbol{w}, \boldsymbol{b}\}$ the distribution of risk (i.e., p values) is described by distribution function $\boldsymbol{F_g}$ with associated density $\boldsymbol{f_g}$. We choose a cut-off level of predicted probability, α, so that 5% of the total population have probabilities are greater than α. That is, α solves:

$$\frac{\boldsymbol{F_w(\alpha)N_w} + \boldsymbol{F_b(\alpha)N_b}}{\boldsymbol{N}} = \boldsymbol{0.95}$$

As explained in the section 2, the tool's designers used a quantile method to assign children into each risk score bin. Specifically, we are interested in the top 5% highest risk group whose LASSO predicted probability is greater than 0.8 (Table 3), or the actual predicted probability of placement is greater than 0.122 (i.e., α = 0.122). Thus $F_w(0.122) \approx \mathbf{0.97}$ and $F_b(0.122) \approx \mathbf{0.89}$.

Thus, an individual with risk level π receives scores such that:

$$S(\pi) = \begin{cases} \boldsymbol{H \; if \; \pi > \alpha} \\ \boldsymbol{L \; if \; \pi \leq \alpha} \end{cases}$$

For a person in group g, the probability

$$\mathbf{Pr}(\boldsymbol{Y = 1 | S = H}) = \frac{\mathbf{1}}{\mathbf{1 - F_g(\alpha)}} \int_{\alpha}^{\mathbf{1}} \boldsymbol{\pi f_g(\pi) d\pi}$$

Since $\begin{cases} \boldsymbol{FPR} = \mathbf{Pr}(\mathbf{S = H | Y = 0}) = \mathbf{1} - \mathbf{Pr}(\boldsymbol{S = L | Y = 0}) = \mathbf{1} - \boldsymbol{TNR} \\ \boldsymbol{FNR} = \mathbf{Pr}(\boldsymbol{S = L | Y = 1}) = \mathbf{1} - \mathbf{Pr}(\boldsymbol{S = H | Y = 1}) = \mathbf{1} - \boldsymbol{TPR} \end{cases}$

we will focus on calculating TNR and TPR.

According to the definition of conditional probability:

$$\mathbf{Pr}(\boldsymbol{S = H \; \& \; Y = 1}) = \mathbf{Pr}(\boldsymbol{S = H | Y = 1}) \times \boldsymbol{P(Y = 1)} = \mathbf{Pr}(\boldsymbol{Y = 1 | S = H}) \times \boldsymbol{P(S = H)}$$

$$\rightarrow \boldsymbol{TPR} = \mathbf{Pr}(\boldsymbol{S = H | Y = 1}) = \frac{\mathbf{Pr}(\boldsymbol{Y = 1 | S = H}) \times \boldsymbol{P(S = H)}}{\boldsymbol{P(Y = 1)}}$$

where,

$$\mathbf{Pr}(\boldsymbol{S = H}) = \mathbf{1} - \boldsymbol{F_g(\alpha)}$$

$$\mathbf{Pr}(\boldsymbol{Y = 1}) = \int_{\mathbf{0}}^{\mathbf{1}} \boldsymbol{\pi f_g(\pi) d\pi}$$

Thus,

$$\boldsymbol{TPR} = \mathbf{Pr}(\boldsymbol{S = H | Y = 1}) = \left(\frac{\mathbf{1}}{\mathbf{1 - F_g(\alpha)}} \int_{\alpha}^{\mathbf{1}} \boldsymbol{\pi f_g(\pi) d\pi} \right) \times \left(\mathbf{1 - F_g(\alpha)} \right) \times \frac{\mathbf{1}}{\int_{\mathbf{0}}^{\mathbf{1}} \boldsymbol{\pi f_g(\pi) d\pi}}$$

$$\rightarrow \boldsymbol{TPR} = \mathbf{Pr}(\boldsymbol{S = H | Y = 1}) = \frac{\int_{\alpha}^{\mathbf{1}} \boldsymbol{\pi f_g(\pi) d\pi}}{\int_{\mathbf{0}}^{\mathbf{1}} \boldsymbol{\pi f_g(\pi) d\pi}}$$

Note that we can use integration by parts to show the following:

$$\int_p^q \boldsymbol{\pi f_g(\pi) d\pi} = \boldsymbol{\pi F_g(\pi)|_{\pi=p}^{\pi=q}} - \int_p^q \boldsymbol{F_g(\pi) d\pi} = \boldsymbol{qF_g(q) - pF_g(p)} - \int_p^q \boldsymbol{F_g(\pi) d\pi}$$

If p = 0 and q = 1, we have:

$$\int_0^1 \pi f_g(\pi) d\pi = 1 - \int_0^1 F_g(\pi) d\pi = \int_0^1 [1 - F_g(\pi)] d\pi$$

This is the area above $\boldsymbol{F_g}$ up to the horizontal line at height 1 in Fig. 5 the sum of the areas indicated by the letters A and B

If p = $\boldsymbol{\alpha}$ and q = 1, we have:

$$\int_\alpha^1 \boldsymbol{\pi f_g(\pi) d\pi} = \boldsymbol{1 - \alpha F_g(\alpha)} - \int_\alpha^1 \boldsymbol{F_g(\pi) d\pi}$$

$$= \boldsymbol{\alpha}\big[\boldsymbol{1 - F_g(\alpha)}\big] + \boldsymbol{(1 - \alpha)} - \int_\alpha^1 \boldsymbol{F_g(\pi) d\pi}$$

$$= \boldsymbol{\alpha}\big[\boldsymbol{1 - F_g(\alpha)}\big] + \int_\alpha^1 \boldsymbol{[1 - F_g(\pi)] d\pi}$$

This is the area above $\boldsymbol{F_g}$ between horizontal lines at heights $\boldsymbol{F_g(\alpha)}$ and 1 indicated by the letter A in Fig. 5 Thus, the TPR for group g is the ratio of the second area over the first: TPR = A/(A+B) hence FNR = B/(A+B)

Similar to the above proof.

$$\boldsymbol{TNR = \Pr(S = L|Y = 0) = \frac{\Pr(Y = 0|S = L) \times P(S = L)}{P(Y = 0)}}$$

$$\boldsymbol{\rightarrow TNR = \Pr(S = L|Y = 0) = \frac{\int_0^\alpha (1 - \pi) f_g(\pi) d\pi}{1 - \int_0^1 \pi f_g(\pi) d\pi}}$$

Note that: $\boldsymbol{1 - \int_0^1 \pi f_g(\pi) d\pi} = 1 - \boldsymbol{\int_0^1 [1 - F_g(\pi)] d\pi} = \boldsymbol{\int_0^1 F_g(\pi) d\pi}$. This is the area under the distribution function $\boldsymbol{F_g}$: the sum of the areas denoted by C and D in Fig. 5. Also, $\boldsymbol{\int_0^\alpha (1 - \pi) f_g(\pi) d\pi} = \boldsymbol{(1 - \alpha) F_g(\alpha)} + \boldsymbol{\int_0^\alpha F_g(\pi) d\pi}$. This is the area below the distribution function up to the horizontal line at height $F_g(\alpha)$, denoted by D in Fig. 5.

Thus, the TNR for group g is the ratio of second area over the first: TNR = D/(C+D) hence FPR = C/(C+D).

Appendix C

Table 1. Distribution of eligible groups by race

	The Priority group	The Family Support	The Universal group	Total
Black	1,448 (57.2%)	4,089 (54.3%)	4,830 (12.7%)	10,367 (21.6%)
White	1,082 (42.8%)	3,439 (42.7%)	33,146 (87.3%)	37,667 (78.4%)
Total	2,530 (100%)	7,528 (100%)	37,976 (100%)	48,034 (100%)

Appendix D

Table 3. Logistic regression on Lasso predicted probabilities for black children[left], and white children[right].

Logit regression output (Black children)	
	Dependent variable:
	Placement by age three
Lasso probabilities	6.7***
	(0.3)
Constant	-7.6***
	(0.3)
Observations	10,367
Log Likelihood	-1,372.5
Akaike Inf. Crit.	2,748.9
Note:	*p<0.1; **p<0.05; ***p<0.01

Logit regression output (White children)	
	Dependent variable:
	Placement by age three
Lasso probabilities	7.4***
	(0.2)
Constant	-7.7***
	(0.2)
Observations	37,667
Log Likelihood	-1,574.5
Akaike Inf. Crit.	3,152.9
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix E

Table 4: Actual and transformed placement risk cut-off

	Actual placement risk cut-off			
Ventile risk score	Lower_bound_actual_probability	Upper_bound_actual_probability	Lower_bound_Lasso_probability	Upper_bound_Lasso_probability
1	0	0.00058049	0.02726574	0.03271670
2	0.00058049	0.00058521	0.03271684	0.03378672
3	0.00058521	0.00058982	0.03378763	0.03473343
4	0.00058982	0.00059585	0.03473370	0.03589946
5	0.00059585	0.00060761	0.03590096	0.03782093
6	0.00060761	0.00062516	0.03782320	0.04144797
7	0.00062516	0.00064974	0.04144803	0.04555632
8	0.00064974	0.00068768	0.04555718	0.05304119
9	0.00068768	0.00079613	0.05304121	0.06927234
10	0.00079613	0.00090163	0.06927611	0.08634506
11	0.00090163	0.00111674	0.08634964	0.11615590
12	0.00111674	0.00154412	0.11619680	0.15579400
13	0.00154412	0.00246244	0.15580640	0.21503670
14	0.00246244	0.00420028	0.21503970	0.28748050
15	0.00420028	0.00754599	0.28749090	0.37158270
16	0.00754599	0.01341331	0.37160700	0.46199920
17	0.01341331	0.02454072	0.46200470	0.54957050
18	0.02454072	0.05221742	0.54960170	0.66423010
19	0.05221742	0.12159590	0.66431800	0.80184890
20	0.12159590	0.41833180	0.80190950	0.99975780

Appendix F

Table 5. Logistic regression on Black variable (i.e., if the child race is black) and Lasso predicted probabilities and interactive terms between Lasso probabilities and race for the black and white sample (N =48,034). Note: We include race in the conversion function from Lasso probabilities to actual abuse risk as we want to see if the estimated true risks differ by race. Since we have other races such as Hispanic, Native American, Asian in the full sample, we want to limit our analysis to black and white children only.

Logit regression output

	Dependent variable:
	Sensitive death outcome by age three
Lasso probabilities	4.2^{***}
	(0.5)
Black	1.8^{***}
	(0.6)
Black*Lasso probabilities	-2.6^{***}
	(0.9)
Constant	-8.4^{***}
	(0.3)
Observations	48,034
Log Likelihood	-466.9
Akaike Inf. Crit.	941.8
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix G

Table 6. Logistic regression on Lasso predicted probabilities for black children[left], and white children[right].

Logit regression output (Black children)	
	Dependent variable:
	Sensitive death outcome by age three
Lasso probabilities	1.6**
	(0.7)
Constant	-6.6***
	(0.4)
Observations	10,367
Log Likelihood	-219.9
Akaike Inf. Crit.	443.8
Note:	*p<0.1; **p<0.05; ***p<0.01

Logit regression output (White children)	
	Dependent variable:
	Sensitive death outcome by age three
Lasso probabilities	4.2***
	(0.5)
Constant	-8.4***
	(0.3)
Observations	37,667
Log Likelihood	-247.0
Akaike Inf. Crit.	498.0
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix H

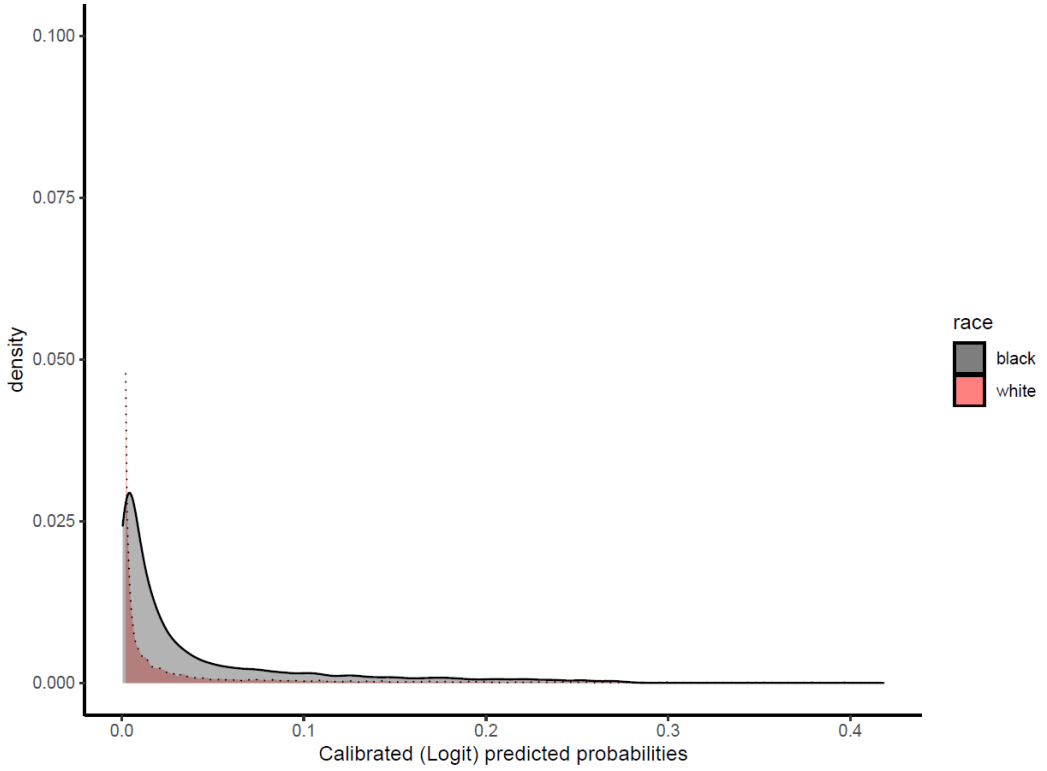

Figure 6: Empirical distribution of estimated true risks by racial group (Density plot).

APPENDIX I

Ethical Approval Letter

15 October 2020

Matthew Ryan

Faculty of Business Economics and Law

Dear Matthew

Re Ethics Application: 20/245 Understanding the mechanism of racial bias in predictive risk models of child welfare.

Thank you for providing evidence as requested, which satisfies the points raised by the Auckland University of Technology Ethics Committee (AUTEC).

Your ethics application has been approved for three years until 15 October 2023.

Standard Conditions of Approval

1. The research is to be undertaken in accordance with the Auckland University of Technology Code of Conduct for Research and as approved by AUTEC in this application.
2. A progress report is due annually on the anniversary of the approval date, using the EA2 form.
3. A final report is due at the expiration of the approval period, or, upon completion of project, using the EA3 form.
4. Any amendments to the project must be approved by AUTEC prior to being implemented. Amendments can be requested using the EA2 form.
5. Any serious or unexpected adverse events must be reported to AUTEC Secretariat as a matter of priority.
6. Any unforeseen events that might affect continued ethical acceptability of the project should also be reported to the AUTEC Secretariat as a matter of priority.
7. It is your responsibility to ensure that the spelling and grammar of documents being provided to participants or external organisations is of a high standard and that all the dates on the documents are updated.

AUTEC grants ethical approval only. You are responsible for obtaining management approval for access for your research from any institution or organisation at which your research is being conducted and you need to meet all ethical, legal, public health, and locality obligations or requirements for the jurisdictions in which the research is being undertaken.

Please quote the application number and title on all future correspondence related to this project.

For any enquiries please contact ethics@aut.ac.nz. The forms mentioned above are available online through http://www.aut.ac.nz/research/researchethics

(This is a computer-generated letter for which no signature is required)
The AUTEC Secretariat
Auckland University of Technology Ethics Committee
Cc: huyen.dinh@aut.ac.nz; Rhema Vaithianathan

References

Ards, S., Myers, S., Malkis, A., Sugrue, E., & Zhou, L. (2003). *Racial Dispropotionality in Reported and Substantiated Child Abuse and Neglect: An Examination of Systematic Bias* (SSRN Scholarly Paper ID 2364533). Social Science Research Network. https://papers.ssrn.com/abstract=2364533

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact Essay. *California Law Review*, *104*(3), 671–732.

Bartholet, E. (2009). The Racial Disproportionality Movement in Child Welfare: False Facts and Dangerous Directions. *Arizona Law Review*, *51*(4), 871–932.

Becker, G. S. (1957). *The economics of discrimination: An economic view of racial discrimination*. University of Chicago.

Cénat, J. M., McIntee, S.-E., Mukunzi, J. N., & Noorishad, P.-G. (2021). Overrepresentation of Black children in the child welfare system: A systematic review to understand and better act. *Children and Youth Services Review*, *120*, 105714. https://doi.org/10.1016/j.childyouth.2020.105714

Chasnoff, I. J., Landress, H. J., & Barrett, M. E. (1990). *The Prevalence of Illicit-Drug or Alcohol Use during Pregnancy and Discrepancies in Mandatory Reporting in Pinellas County, Florida*. https://www.nejm.org/doi/full/10.1056/NEJM199004263221706

Chibnall, S., Dutch, N. M., Jones-Harden, B., Brown, A., & Gourdine, R. (2003). *Children of Color in the Child Welfare System: Perspectives from the Child Welfare Community*. 102.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Chouldechova, A., & Roth, A. (2018). The Frontiers of Fairness in Machine Learning. *ArXiv:1810.08810 [Cs, Stat]*. http://arxiv.org/abs/1810.08810

Clarke, J. (2011). The challenges of child welfare involvement for Afro-Caribbean families in Toronto. *Children and Youth Services Review*, *33*(2), 274–283. https://doi.org/10.1016/j.childyouth.2010.09.010

Cowgill, B., & Tucker, C. E. (2019). Economics, Fairness and Algorithmic Bias. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3361280

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R., & Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review*, *79*, 291–298. https://doi.org/10.1016/j.childyouth.2017.06.027

Dalgleish, L. I. (2006). Testing for the effects of decision bias on overrepresentation: Applying the GADM model. *16th International Congress on Child Abuse and Neglect*.

Dettlaff, A. J., Rivaux, S. L., Baumann, D. J., Fluke, J. D., Rycraft, J. R., & James, J. (2011). 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. https://doi.org/10.1016/j.childyouth.2011.04.005

Drake, B., Jonson-Reid, M., Ocampo, M. G., Morrison, M., & Dvalishvili, D. (Daji). (2020). A Practical Framework for Considering the Use of Predictive Risk Modeling in Child Welfare. *The ANNALS of the American Academy of Political and Social Science*, *692*(1), 162–181. https://doi.org/10.1177/0002716220978200

Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., Koss, M. P., & Marks, J. S. (1998). Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study. *American Journal of Preventive Medicine*, *14*(4), 245–258. https://doi.org/10.1016/S0749-3797(98)00017-8

Fluke, J. D., Yuan, Y.-Y. T., Hedderson, J., & Curtis, P. A. (2003). Disproportionate representation of race and ethnicity in child maltreatment: Investigation and

victimization. *Children and Youth Services Review*, *25*(5–6), 359–373. https://doi.org/10.1016/S0190-7409(03)00026-4

Font, S. A., Berger, L. M., & Slack, K. S. (2012). Examining Racial Disproportionality in Child Protective Services Case Decisions. *Children and Youth Services Review*, *34*(11), 2188–2200. https://doi.org/10.1016/j.childyouth.2012.07.012

Gillingham, P. (2016). Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning. *British Journal of Social Work*, *46*(4), 1044–1058. https://doi.org/10.1093/bjsw/bcv031

Glaberson, S. K. (2019). Coding Over the Cracks: Predictive Analytics and Child Protection Artificial Intelligence and Predictive Algorithms: Why Big Data Can Lead to Big Problems. *Fordham Urban Law Journal*, *46*(2), 307–363.

Goldhaber-Fiebert, J. D., & Prince, L. (2019). *Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County's Child Welfare Office*. 97.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Harris, M. S., & Hackett, W. (2008). Decision points in child welfare: An action research model to address disproportionality. *Children and Youth Services Review*, *30*(2), 199–215. https://doi.org/10.1016/j.childyouth.2007.09.006

Hastie, T., Qian, J., & Tay, K. (2021). *An Introduction to glmnet*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Linear Model Selection and Regularization. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: With Applications in R* (pp. 203–264). Springer. https://doi.org/10.1007/978-1-4614-7138-7_6

King, B., Fallon, B., Boyd, R., Black, T., Antwi-Boasiako, K., & O'Connor, C. (2017). Factors associated with racial differences in child welfare investigative decision-making in Ontario, Canada. *Child Abuse & Neglect*, *73*, 89–105. https://doi.org/10.1016/j.chiabu.2017.09.027

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, *105*(5), 491–495. https://doi.org/10.1257/aer.p20151023

Knowles, J., Persico, N., & Todd, P. (2015). Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy*. https://doi.org/10.1086/318603

Lane, W. G. (2002). Racial Differences in the Evaluation of Pediatric Fractures for Physical Abuse. *JAMA*, *288*(13), 1603. https://doi.org/10.1001/jama.288.13.1603

Lavergne, C., Dufour, S., Trocmé, N., & Larrivée, M.-C. (2008). Visible Minority, Aboriginal, and Caucasian Children Investigated by Canadian Protective Services. *Child Welfare*, *87*, 59–76.

Lu, Y. E., Landsverk, J., Ellis-Macleod, E., Newton, R., Ganger, W., & Johnson, I. (2004). Race, ethnicity, and case outcomes in child protective services. *Children and Youth Services Review*, *26*(5), 447–461. https://doi.org/10.1016/j.childyouth.2004.02.002

McCallum, K., & Cheng, A.-L. (2016). *Community Factors in Differential Responses of Child Protective Services*. https://doi.org/10.1111/phn.12214

Miller, M. G. (2008). *Racial Disproportionality in Washington State's Child Welfare System*.

Mullainathan, S., & Obermeyer, Z. (2017). Does Machine Learning Automate Moral Hazard and Error? *American Economic Review*, *107*(5), 476–480. https://doi.org/10.1257/aer.p20171084

Mullainathan, S., & Obermeyer, Z. (2021). On the Inequity of Predicting A While Hoping for B. *AEA Papers and Proceedings*, *111*, 37–42. https://doi.org/10.1257/pandp.20211078

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Passi, S., & Barocas, S. (2019). Problem Formulation and Fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. https://doi.org/10.1145/3287560.3287567

Pessach, D., & Shmueli, E. (2020). Algorithmic Fairness. *ArXiv:2001.09784 [Cs, Stat]*. http://arxiv.org/abs/2001.09784

Rivaux, S., James, J., Wittenstrom, K., Baumann, D., Sheets, J., Henry, J., & Jeffries, V. (2008). The Intersection of Race, Poverty, and Risk: Understanding the Decision to Provide Services to Clients and to Remove Children. *Child Welfare*, *87*, 151–168.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vaithianathan, R., Benavides-Pradp, D., & Putnam-Hornstein, E. (2020). *Implementing the Hello Baby Prevention Program in Allegheny County*. https://espace.library.uq.edu.au/view/UQ:f7909e0