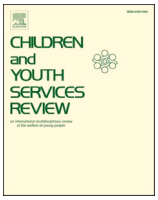
Children and Youth Services Review 145 (2023) 106777

Contents lists available at ScienceDirect

Children and Youth Services Review

journal homepage: www.elsevier.com/locate/childyouth



The pursuit of algorithmic fairness: On "Correcting" algorithmic unfairness in a child welfare reunification success classifier

Jordan Purdy *, Brian Glass

Office of Reporting Research Analytics and Implementation, Oregon Department of Human Services, United States

A B S T R A C T

The algorithmic fairness of predictive analytic tools in the public sector has increasingly become a topic of rigorous exploration. While instruments pertaining to criminal recidivism and academic admissions, for example, have garnered much attention, the predictive instruments of Child Welfare jurisdictions have received considerably less attention. This is in part because comparatively few such instruments exist and because even fewer have been scrutinized through the lens of algorithmic fairness. In this work, we address both gaps. First, a novel classification algorithm for predicting reunification success within Oregon Child Welfare is presented. The purpose of this tool is to maximize the number of stable reunifications and identify potentially unstable reunifications which may require additional resources and scrutiny. Additionally, because the algorithmic fairness of the developed tool, if left unaltered, is unquestionably lacking, the utilized procedure for mitigating such unfairness is presented, along with the nuanced rationale behind each complex and unavoidable choice. This procedure, though similar to other post-processing group-specific thresholding methods, is novel in its use of a penalized optimizer and contextually requisite subsampling. These novel methodological components yield a rich and informative empirical understanding of the trade-off continuum between fairness and accuracy. As the developed procedure is generalizable across a variety of group-level definitions of algorithmic fairness, as well as across an arbitrary number of protected attribute levels and risk thresholds, this approach presents the opportunity to critically and broadly influence the equity and fairness implications of a Child Welfare jurisdiction's automated decision processes.

1. Introduction

The Child Welfare division of the Oregon Department of Human Services operates under a mission to "strengthen, preserve, and reunify families" (Oregon Department of Human Services, 2015), as well as a mission to "adapt services and policy to eliminate discrimination and disparities in the delivery of human services" (Oregon Department of Human Services, 2014). In order to advance both principles in practice, we present a novel methodology which identifies children in custody of the state who are candidates for stable reunification with their family, and which "corrects" algorithmic unfairness in the corresponding automated classification process. The intended implementation of this process is in the form of a decision support tool for staff who make permanency-related decisions in Child Welfare.

The purpose of such a decision support tool is to advance the mission to safely reunify children with their families by identifying the probability of a failed reunification. Currently, of children who have been removed from home and placed in substitute care for at least 90 days, only 36% will reunify with their family within one year. However, 83% of reunifications remain stable for at least one year. Taken together, we seek to develop a classification procedure to ascertain the probability of a stable reunification in order to (1) maximize the amount and success rate of reunifications, and (2) identify high risk reunifications which may benefit from supportive services and resources (Table 1).

The overall reunification rate is the result of a variety of competing factors such as policy and practice models (Ainsworth & Maluccio, 1998), social worker decision making (Biehal, Sinclair, & Wade, 2015), jurisdiction resources (Esposito et al., 2017), and parent/child characteristics and behavior (Biehal, 2007; Terling, 1999). In this way, decision support tools face systemic challenges beyond algorithmic performance and fairness (Keddell, 2019). However, the methodology presented here represents a general child-level predictive risk analysis. The output of such an analysis has a variety of potential applications, from a system-wide evaluation of a jurisdiction's reunification practice, to informing targeted service or resource initiatives, to real-time individual-level decision support tools. The correction process detailed here is not limited to decision support tools, but rather represents a general process for addressing algorithmic bias in individual-level predictive risk assessment.

The construction of the proposed equitable classifier involves two stages, resulting in the assessment of an individual child's prospects for a stable reunification along an ordinal risk tier scoring system. First, a

* Corresponding author at: Oregon Department of Human Services, 500 Summer St. NE, Salem, OR 97302, United States.
E-mail addresses: jordan.e.purdy@dhsoha.state.or.us (J. Purdy), brian.d.glass@dhsoha.state.or.us (B. Glass).

https://doi.org/10.1016/j.childyouth.2022.106777
Received 21 April 2021; Received in revised form 3 June 2022; Accepted 20 December 2022
Available online 23 December 2022
0190-7409/Published by Elsevier Ltd.

Table 1
Calibration table for risk score outcomes and proportions for two groups: (1) children who leave substitute care to reunify with their parents, and (2) children who have not yet reunified with their parents at 90 days from entry into substitute care.

Score	Reunifications			In Care at 90 Days		
	% GivenScore	% Failed Reunification		% GivenScore	% Eventually Reunify	
		w/in 1 Year	Ever		w/in 1 Year	w/in 1 Year, then Fail
S4	8%	52%	60%	8%	31%	30%
S3	26%	25%	35%	31%	33%	22%
S2	34%	12%	21%	35%	35%	14%
S1	32%	7%	14%	26%	40%	10%
Overall	100%	17%	25%	100%	36%	16%

binary classifier is constructed using statistical machine learning. Such a classifier, as far as we are aware, is the first of its kind in Child Welfare at large. Second, because such a classifier may suffer from myriad forms of bias, we administer a post-processing fairness correction. This procedure adjusts the binary classification threshold dependent on the child's level of their protected attribute (e.g., demographic group membership). By combining multiple binary classification thresholds, a multi-tiered ordinal scoring system can be constructed. Here, a four-score system is developed using three thresholds.

As child welfare agencies continue to turn to predictive risk modeling to support human decision making, there are many overarching issues these agencies must consider. The present work considers a method for addressing the algorithmic bias generated or perpetuated by such modeling, and not a general discussion of the benefits or detriments of predictive risk modeling in the first place (Drake, Jonson-Reid, Ocampo, Morrison, & Dvalishvili, 2020). Once deployed, these automated algorithms may impact every child coming into contact with a jurisdiction's child welfare agency, and at multiple time points through a child's developmental process. For this reason, the statistical mechanisms presented in this article have the opportunity for critical and widespread influence of the equity and fairness implications of a jurisdiction's decision process.

1.1. Making space for algorithmic fairness

Machine learning classifiers have become widespread in the private sector (Einav & Levin, 2014) and their use is expanding in the public domain (Oswald, 2018). There exist serious and well founded concerns over the inherent unfairness or bias in these algorithms. Unfairness may be introduced into machine learning classifiers from multiple sources: externally from the historical decision-making processes which generated the training data, or internally from statistical artifacts themselves (Veale, Van Kleek, & Binns, 2018).

In an effort to address these concerns, two approaches are not uncommonly suggested:

1. Remove the protected attribute from the feature set during algorithm training.
2. Verify the protected attribute is not statistically significant in the model.

Such methods are ultimately misguided conceptions of what constitutes algorithmic fairness within a binary classification task. Both are akin to a "fairness through unawareness" approach, but both ultimately fail to acknowledge the likely relationships that exist between the protected attribute and the other features accessible to the algorithm. Berk et. al. phrase it this way: "Even when direct indicators of protected group membership, such as race and gender, are not included as predictors, associations between these measures and legitimate predictors can "bake in" unfairness" (Berk, Heidari, Jabbari, Kearns, & Roth, 2017, pg. 2). Hardt et. al., referencing (Pedreshi, Ruggieri, & Turini, 2008), convey it this way: "…this idea of 'fairness through unawareness' is ineffective due to the existence of *redundant encodings*, [or] ways of predicting protected attributes from other features" (Hardt, Price, & Srebro, 2016, pg. 1). In particular, the first conception problematically has not removed from the feature set any proxies for the protected attribute, while the second has problematically only verified that the protected attribute is not significant *after accounting for the effects of all other variables in the model*. Hence, algorithmic fairness is not achieved by either (only) removing the protected attribute from the feature set or by only "verifying" its lack of statistical significance.

In this work we present and utilize a post-processing fairness correction technique which seeks to address unfairness of the final automated classifier, regardless of the initial sources of the unfairness. This procedure, though similar to other post-processing group-specific thresholding methods, is novel in its unexpected use of a penalized optimizer, its contextually necessary use of subsampling, and its generalizability across any of nine group-level definitions of algorithmic fairness. Through such novelty, the procedure yields a rich and informative empirical understanding of the trade-off continuum between fairness and accuracy. Given that the procedure can also accommodate an arbitrary number of protected attribute levels and an arbitrary number of risk thresholds, the approach is broadly applicable both within and beyond Child Welfare.

This proactive approach to mitigating a lack of algorithmic fairness represents a substantial departure from the current norm among the growing number of Child Welfare jurisdictions developing and deploying predictive risk tools; see Samant, Horowitz, Xu, and Beiers (2021) for a comprehensive detailing of all such jurisdictions. Excluding Oregon (Office of Reporting Research Analytics & Implementation, 2019; Purdy, Glass, & Pakseresht, 2018), the current "standard" is either to essentially ignore algorithmic fairness altogether or to perform an assessment-only audit (e.g., Chouldechova, Putnam-Hornstein, Benavides-Prado, Fialko, & Vaithianathan (2018b)). In fact, the only application of a fairness correction procedure to a predictive analytic tool in Child Welfare, apart from the tools developed by ORRAI, is academic and illustrative in nature (i.e., Coston, Mishler, Kennedy, & Chouldechova, 2020), similar to the number of academic papers pertaining to predictive policing and criminal recidivism (e.g., Canetti et al., 2019; Ensign, Friedler, Neville, Scheidegger, & Venkatasubramanian, 2018) that can be linked to Angwin, Larson, Mattu, and Kirchner (2016). This de facto state of affairs in Child Welfare is undoubtedly attributable, in no small part, to the uncertainty surrounding how "best" to proceed.

To address this uncertainty, we provide the rationale behind each critical decision embedded in the procedural methodology. We hope such transparency serves as a "map" for other jurisdictions to follow and, where appropriate, deviate from accordingly. In the case of the reunification algorithm, critical decisions were necessarily informed and influenced by the stakeholders comprising the algorithm's work group, in accordance with their knowledge of and desire for Child Welfare within the state of Oregon. Listed in alphabetical order, this work group consisted of the following staff and representatives: business analysts, Child Protective Services (CPS) supervisors, CPS workers, child safety manager, Child Welfare (CW) alcohol and drug specialist, CW district managers, CW field leadership, CW leadership, current and former foster youth, current and former foster parents, data coordinator, Indian Child Welfare Act representative from Tribal Unit, lead inter-agency researcher, Mentoring Assisting Promoting Success (MAPS) worker, Morison Child and Family Services, Office of Equity and Multicultural Services, Office of Information Services, ORRAI reporting analysts, Oregon Department of Justice, paralegals, permanency consultant, permanency program manager, permanency supervisors, permanency workers, program managers, program systems support, reunification manager, safety consultant, supervisor, and teen supervisor. Ultimately, the presented procedure can and surely will be improved upon over

time, but we hope that it serves as a new "standard" for those Child Welfare jurisdictions seeking to proactively address algorithmic fairness.

1.2. Structure of paper

In Section 2, details of the algorithm, including its development and application, are provided. In Section 3, the identified protected attribute, the chosen definition of algorithmic fairness, and the developed methodology for fairness "correcting" are described. The results of the fairness correction procedure, when applied to the reunification algorithm, are detailed in Section 4. The paper concludes in Section 5 with a discussion of opportunities for future exploration. To broadly encourage the incorporation of algorithmic fairness into Child Welfare decision support systems, and to freely provide a means of mitigating a lack of algorithm fairness in such systems, the R code for the developed fairness correction procedure is posted on GitHub (https://github.com/JPurdy-ORRAI/ORRAI_AlgFairnessCorrectionDemo), along with an illustrative R script applying the procedure to the Adult Data Set from the UCI Machine Learning Repository (Dheeru & Karra Taniskidou, 2017).

2. Predicting reunification success

2.1. Data sources, variables, and outcome

The primary data source was an administrative data set queried from the state of Oregon's Child Welfare data system. The data system is compliant to the U.S. Children's Bureau's requirements for a Statewide Automated Child Welfare Information System (SACWIS) (DHHS, 2016).

The general unit of observation was a child's transition from one placement setting to another (i.e., a child-transition pair). The machine learning training set consisted of only child-reunification pairs, defined as the child-transition pairs representing a return of the child from a substitute care setting to a home setting with the child's parents. These reunification observations consisted of both "trial visits" (i.e., temporary reunifications during which the child remains in state custody) and full reunifications (i.e., discharge from state custody, with or without ongoing government provided support services). The data set queried from this data source represented child welfare administrative data from August 2011 to January 2020.

The outcome of interest was whether a child's potential reunification with their family will be stable. To quantitatively define this outcome (Passi & Barocas, 2019), the binary dependent variable was deemed "true" if the child remained at home for a period of one year, and "false" if the child returned to substitute care for at least 14 days during the ensuing one-year period after returning home. Most children who have experienced at least one reunification event have experienced multiple ($25^{th}percentile = 1$, $M = 2.7$, $75^{th}percentile = 3$). After stochastically unduplicating by child (i.e., choosing one reunification event per child), the overall outcome prevalence for a reunification failure was 17%.

The independent variables (i.e., the machine learning feature set), were constructed from data elements which would have been temporally available on the day before a child-reunification pair occurred. In this way, temporal leakage was prevented by ensuring the machine learning classifier could not "peek" at information about the child's new setting nor about the child's future administrative data pattern. Features were only constructed using timestamped data elements with consistent data entry availability throughout the life of the SACWIS system. The features were constructed using information regarding prior: service placements in substitute care, home-based government provided child welfare service involvement, reports of abuse/neglect, and CPS investigations. The features convey information related not only to the child, but also the child's parents and the perpetrator listed on the child's most recent CPS investigation. The complete list of features is available in Table 9 in Appendix D.

2.2. From machine learning classifier to decision support tool

Careful procedural protocols are required to avoid introducing artifacts via the modeling procedure itself. These "modeling pitfalls" include the selective label problem, repeated observation or temporal leakage, data volume as predictor, shrinking outcome windows, and inappropriate performance metrics. Each are discussed in Appendix A.

A machine learning classifier is trained to predict the outcome with the available child-reunification-level feature set. The classifier was trained using gradient-boosted decision trees via the XGBoost algorithm (Chen, He, Benesty, Khotilovich, & Tang, 2015). Because SACWIS system data are not conducive to exploring or understanding the causal mechanisms underlying reunification failure or success, we used a decision tree classifier to maximize predictive performance through the leveraging of complex interactions between variables (Breiman, 2001). The classifier training procedure involves an exhaustive resampling approach to ensure different events involving the same child do not appear in both the training and test data sets. The procedure yields an independent test-set-predicted-probability of reunification stability for each child-reunification observation in the complete set (see: Appendix A.2).

To conform to the design and implementation specifications of Oregon's Child Welfare governance, the classifier's output was adapted for use as a decision support tool. To accomplish this, three risk thresholds were selected to represent usable and meaningful risk tiers for practitioners, resulting in four ordinal risk score tiers. To facilitate intuitive understanding among tool users, the threshold (prior to applying the fairness-correction procedure) separating scores of *S1* and *S2* from scores of *S3* and *S4* is the average predicted probability outputted by the algorithm. This enables staff using the tool to quickly identify children with scores of *S1* and *S2* as having less than the "typical" (i.e., average) risk of a failed reunification, and children with scores of *S3* and *S4* as having more than the "typical" (i.e., average) risk. The threshold separating a score of *S1* from higher scores is the median of the predicted probabilities less than the average predicted probability. This provides staff with two equal-sized "bins" in which scores of *S1* convey approximately half the risk of scores of *S2*. The threshold separating a score of *S4* from lower scores represents the 75^{th}-percentile of predicted probabilities greater than the average predicted probability. This choice simultaneously limits the proportion of child-reunification pairs receiving a score of *S4* and maintains an approximate doubling of risk with each subsequent increase in risk score. These thresholds are labeled low-, average-, and high-risk, while the corresponding four risk score tiers are denoted *S1*, *S2*, *S3*, and *S4*, where, for example, the low-risk threshold separates risk scores of *S1* from risk scores of *S2*, *S3*, and *S4*. Such coarse grained risk tier information seeks to reduce the introduction of bias via variable and uncontrollable decision thresholds which can vary between human decision makers (Chouldechova, Benavides-Prado, Fialko, & Vaithianathan, 2018a; Green & Chen, 2019). Finally, an extended data set of observations were constructed to consider performance generalization to all child-placement observations 90 days from initial entry into substitute care (the 90 day mark was chosen to mirror the intended implementation protocol for this decision support use case). This generalization process was required due to the fact that only historical reunificaitons could be directly compared to model results. In other words, by definition there are no outcomes available for children who did not experience a reunification. Thus, in order to assess the implications of the decision support tool, predictive performance was tested on a generalized set using proxy outcomes.

2.3. Predictive performance

To assess predictive performance, Table 1 reports the model's predictive characteristics for child-reunification pairs as well as for children who remained in substitute care at 90 days after entry. As the classifier was trained using child-reunification pairs, the Reunification group

represents a direct predictive test of the classifier. Table 1 indicates that of the Reunifications with a high-risk score (i.e., *S4*), 52% resulted in a reunification failure (i.e., a return to substitute care). In contrast, of those with a low-risk score (i.e., *S1*), 7% resulted in a failure. The column *%Failed Reunification Ever* relaxes the outcome window censorship to illustrate the robustness of the risk stratification to an open-ended outcome.

The In Care group represents a potential use case of the classifier to assess children in substitute care at 90 days from entry. Given the impossibility to observe the reunification failure rate for children who were not historically reunified, a group of children were considered who were In Care and may have eventually reunified. Thus, this In Care group is a generalized use of the classifier, in that it represents an observation type that is not used in the training of the classifier. Note that 36% of the children in the In Care group experience a reunification event in the upcoming year. This proportion represents the critical opportunity to provide a decision support tool to advance the mission to reunify children with their families. In particular, from Table 1, it is apparent that among the 27% of children who have been in care for 90 days and would receive a low-risk score (i.e., *S1*), only 40% will experience an attempted reunification in the next year; whereas, among the 8% of children who have been in care for 90 days and would receive a high-risk score (i.e., *S4*), a comparatively high 31% will experience an attempted reunification in the next year. The lack of clear separation in these and the other values of the *%Eventually Reunify w/in 1 Year* column of Table 1 suggest that the diagnosticity of the tool could greatly enhance reunification-related decisions in Oregon Child Welfare. Moreover, the rightmost column of Table 1 illustrates that the reunification failure rate of the group of children who did reunify was calibrated with scoring, indicating that the classifier has the potential to provide additional outcome-based insights to those responsible for reunification decisions.

The area under the receiver operating characteristic curve (AUC/ROC) was 0.73. Despite this moderate AUC value, the low historical reunification rate coupled with the lack of diagnosticity of the historical reunification decisions represents a real opportunity to support human decision making in this area. The scores provided in Table 1 are generated using the fairness corrected version of the classifier. The fairness correction procedure is discussed below.

3. The algorithmic fairness of the reunification scores

A general framework for enmeshing algorithmic fairness in a binary decision-support tool, both within and beyond Child Welfare, requires answering three questions. First, what is the protected attribute across which fairness is assessed? Second, what is the definition through which fairness is measured? Third and finally, what is the procedure from which fairness is increased?

Before providing and discussing the answers to these three questions within the context of the reunification tool, it is important to consider who is responsible, in general, for answering these three questions. In our opinion, providing answers to the first two questions is the responsibility of the stakeholders of the corresponding decision point. In particular, because all available answers yield unavoidable trade-offs and because the "best" answers are both use-case- and jurisdiction-dependent, the values and worldviews represented in these answers should be driven by stakeholders. Providing an answer to the third question, in our opinion, is the responsibility of the algorithm-development team. In particular, after accounting for the answers provided to the first two questions, the developers are tasked with identifying and implementing a procedure that will maximally increase fairness, subject to as minimal a decrease in predictive performance as the algorithm's stakeholders are willing to accept.

3.1. The protected attribute

In answer to question one, the stakeholders within the reunification algorithm work group identified a race and ethnicity-based protected attribute. Given that algorithms can perform differently across different combinations of protected attributes, as demonstrated in Buolamwini and Gebru (2018), a multi-dimensional protected attribute would have ideally been constructed with a level for each combination of levels across all available protected attributes (i.e., race, ethnicity, ICWA-status, and sex). Unfortunately, sample sizes were insufficiently large to enable such a multidimensional protected attribute, and the work group was forced to make an unavoidable trade-off. The resultant feature consists of four levels: Black (BL); Hispanic, Pacific Islander, or Asian (HPA); Native American or Indian Child Welfare Act (ICWA)-eligible (NV); and White (WH). Importantly, BL and NV child-reunification pairs are historically approximately 47% and 39% more likely, on average, to experience a failed reunification than HPA child-reunification pairs, and historically approximately 20% and 14% more likely, on average, to experience the adverse event than WH child-reunification pairs. It is this disproportionality in historical decision making, along with the ways that it may manifest itself through the predictions of the algorithm, that necessitate a specific answer to the second question.

3.2. The definition of algorithmic fairness

In answer to question two, the stakeholders within the reunification algorithm work group identified Error Rate Balance as the definition through which to measure fairness. The decision to utilize a group-level definition of algorithmic fairness, as opposed to a causal-reasoning based or individual-level definition, was pragmatic in nature. In particular, given that it is fundamentally "…impossible to test an existing classifier against causal definitions of fairness" (Verma & Rubin, 2018, pg. 6), all causal-reasoning based definitions were sweepingly dismissed. Similarly, given that individual-level definitions "…currently cannot be operationalized in a useful manner" (Berk et al., 2017, pg. 15), and given that the worldview represented through any such definition likely has a commensurate analogue within the existing set of operationalizeable group-level definitions (Binns, 2019), individual-level definitions were also sweepingly dismissed.

The decision to use Error Rate Balance from among all available group-level definitions, was ultimately rooted in the worldviews it represents. From a technical perspective, Error Rate Balance requires, at each threshold, that the false positive rate be the same across the levels of the protected attribute, as well as the false negative rate. Contextually, this means that given an outcome for a child-reunification pair, be it "success" or "failure", Error Rate Balance requires that the probability of a corresponding incorrect prediction label be the same across all levels of the protected attribute. More practically, this means Error Rate Balance "…encourages the use of features that allow to directly predict [the outcome], but prohibits abusing [the protected attribute] as a proxy for [the outcome]" (Hardt et al., 2016, pg. 3). Philosophically then, Error Rate Balance "…aims to account for Population Inequity: it strives for risk predictions that do not disproportionately harm one group more than another, regardless of the underlying distributions of risk" (Green, 2020, pg. 8).

This choice of definition, however, does come with a cost. Specifically, the proportion of child-reunification pairs within each risk score that actually end up experiencing the adverse event will differ across protected attribute levels, which can "…have the unintended and highly undesirable consequence of incentivizing [tool users] to take [the protected attribute] into account when interpreting predictions" (Pleiss, Raghavan, Wu, Kleinberg, & Weinberger, 2017, pg. 1). In fact, the choice of any definition comes at the cost of other fairness definitions. Collectively, such trade-offs are known as Impossibility Theorems (Berk et al., 2017), a number of which have been documented in the literature

(Barocas, Hardt, & Narayanan, 2018; Chouldechova, 2017; Ensign et al., 2018; Kleinberg, Mullainathan, & Raghavan, 2016). Ultimately, such costs must be considered and weighed by stakeholders when choosing a definition of fairness and then, where possible, effectively addressed and managed in training users of the corresponding tool.

3.3. The procedure for increasing fairness

In answer to question three, a protected-attribute-level-specific (i.e., group-specific) thresholding adjustment procedure was utilized. In the context of the reunification algorithm, such an approach means that the amount of evidence (i.e., the value of the predicted probability) required to elevate the level of risk (e.g., from a score of *S2* to *S3*) for a child-reunification pair depends on the level of the protected attribute. Furthermore, the "dial" for determining the "appropriate" amount of evidence for each protected attribute level is "tuned" according to the specified definition of algorithmic fairness, which in this application is Error Rate Balance.

While this procedure attains algorithmically fairer risk scores through the same mechanism (i.e., group-specific threshold values) as the procedures described in Hardt et al. (2016) and Lipton, Chouldechova, and McAuley (2019), the optimization process for identifying the values for these thresholds is notably different and, to the best of our knowledge, novel in its application. This optimization process simultaneously accounts for the dependence structure embedded in the observational units (i.e., dependent child-reunification pairs), yields an empirically derived curve of the trade-off continuum between fairness and accuracy, and accommodates any of nine group-level definitions of algorithmic fairness. When combined with the fact that this procedure is applicable to any feasibly chosen protected attribute, the procedure represents an important contribution to the body of literature on algorithmic fairness in the predictive risk tools of Child Welfare.

The decision to utilize such a post-processing procedure, as opposed to a pre- or in-processing procedure similar to those described in Berk et al. (2017), Bechavod and Ligett (2017), or Zafar, Valera, Rodriguez, and Gummadi (2017), was attributable to self-imposed demands for transparency, interpretability, and operationalizability. In particular, this procedure operates outside of the "black-box" of the algorithm so that there is no "mystery" behind the algorithmic fairness process. Such transparency allows stakeholders to straightforwardly recognize how unfairness is being mitigated. When it comes to interpretability, this procedure enables a clear answer to the following question: How much did the Error Rate Balance increase, and corresponding predictive performance measures change, in transitioning from group-agnostic to group-specific values at each threshold? Such interpretability enables stakeholders to clearly recognize the impact of the procedure and to begin to assess the associated trade-offs available. Finally, with respect to operationalizability, this procedure is flexible and transferable across use-cases and classifier types, enabling its broad application.

3.3.1. Clarifying point

It is worth addressing here a common and instinctive rhetorical question to such a process: But is it not unfair to uphold different "standards" for different protected attribute levels when assigning risk scores? Such a question is, at its core, pushing back against the notion that to prevent disparate impact requires disparate treatment, and is in fact the motivation behind the in-processing approach proposed in Zafar et al. (2017). In reality, however, such approaches seeking to prevent disparate impact without disparate treatment (1) fail to optimally prevent disparate impact and (2) ultimately do enact disparate treatment "…through hidden changes to the learning algorithm" (Lipton et al., 2019, pg. 16). Hence, such a criticism is not limited to a group-specific thresholds approach, but in fact applies to a broad range of fairness correction procedures. And in response to this more general criticism, we point out that such disparate treatment is in fact enacting a form of service equity, a core value of the Oregon Department of Human Services.

3.3.2. Brief overview of the procedure

The developed procedure ultimately utilizes a penalized optimizer within a subsampling loop to obtain the corresponding group-specific threshold values. The subsampling component of this procedure ensures that no two reunifications for the same child are simultaneously used in the objective function, which addresses concerns surrounding (1) the dependence structure of the observational units, and (2) the robustness of fairness-correction procedures to training-test splits (Friedler et al., 2019). To then quantify the extent to which Error Rate Balance is achieved for a given subsample of child-reunification pairs, we identify at each threshold the greatest disparity in either false negative rates or false positive rates across any pairing of protected attribute levels. This value, at each threshold, is calculated such that it ranges continuously between 0 and 1, with larger values indicating greater similarity in false positive rates and false negative rates (i.e., greater Error Rate Balance). For example, a value of 0.5 indicates that at least one level of the protected attribute has an error rate that is half that of some other protected attribute level at the specified threshold.

The penalized objective function of this procedure then finds the group specific threshold values that minimize the amount of unfairness within a given subsample, subject to the constraint that the accuracy costs incurred by shifting from group-agnostic threshold values to group-specific threshold values do not exceed a bounded amount. This bounded amount is functionally achieved through a gated weighting mechanism that ranges continuously between zero and one. When this weight is set equal (i.e., swiveled) to zero, the optimization process will yield the fairest group-specific threshold values for the given subsample without any regard for accuracy costs. When this weight is set equal (i.e., swiveled) to one, the optimization process is exclusively concerned with accuracy costs and, consequently, the returned group-specific threshold values for the given subsample will be identical to the group-agnostic threshold value. More generally, as the value of the weight is increased over the range of values between these extremes (i.e., swiveled closer and closer to one), the pursuit of fairness becomes more heavily anchored to the accuracy achieved through the group-agnostic threshold value. By iterating the procedure across a tuning grid for this weight, ranging from zero to one, the functional relationship between accuracy and fairness can be empirically derived at each threshold.

To be clear, the accuracy cost in this penalized objective function is not a direct measure of change to any specific type of predictive accuracy, but is instead the proportion of changing risk scores (e.g., *S*2 changes to *S*3 or *S*3 changes to *S*2), which serves as a broad catch-all for changes across any predictive performance measures. In particular, when a risk score changes as a result of transitioning from the group-agnostic threshold value to the group-specific threshold values, all predictive performance measures are unavoidably impacted, though not necessarily for the worse. Hence, the greater the proportion of changing risk scores, the greater the potential accuracy-related costs.

By then applying this procedure across a large number of random subsamples (e.g., 200) and averaging the resultant group-agnostic and group-specific threshold values across these subsamples, the corresponding pre- and post-fairness corrected threshold values at a specified gated penalty weight are obtained. The entirety of the procedure is provided in Appendix B, along with additional details regarding the quantification of Error Rate Balance and any of the other eight group-level definitions of fairness, as given in Verma and Rubin (2018), that can be utilized with the procedure: Calibration, Conditional Use Accuracy Equality, Equalized Odds, Overall Accuracy Equality, Predictive Equality, Predictive Parity, Statistical Parity, and Treatment Equality.

3.4. Applicability to non-machine-learning decision systems

While this section focused on the approach to algorithmic fairness

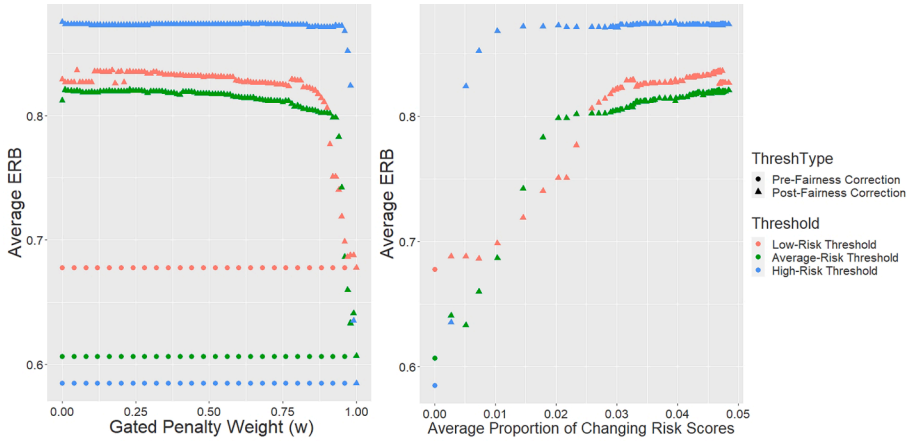


Fig. 1. For each of the low-, average-, and high-risk thresholds, for both the pre- and post-fairness corrected threshold values, average Error Rate Balance versus gated penalty weight is plotted in the left-hand graphic, while average Error Rate Balance versus the average proportion of changing risk scores (i.e., the empirically derived trade-off continuum between fairness and accuracy) is plotted in the right-hand graphic.

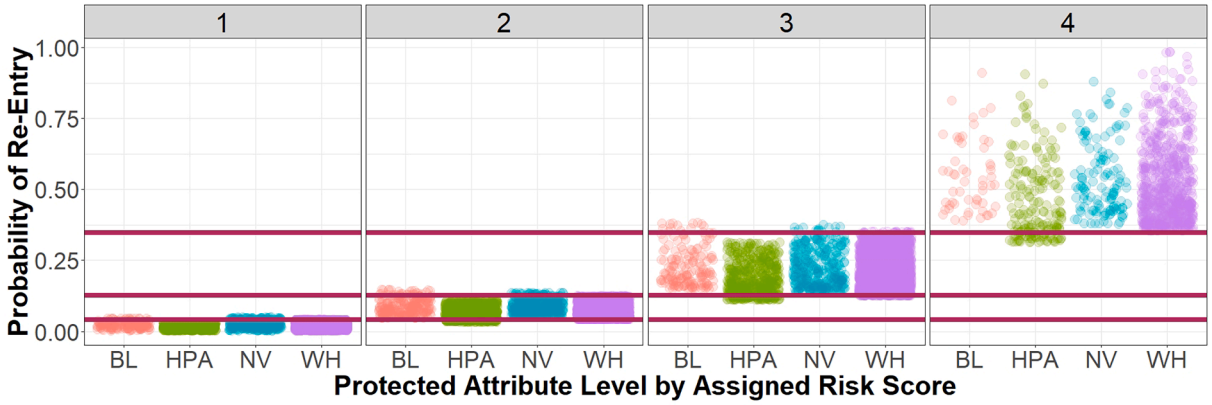


Fig. 2. The predicted probability for each child-reunification pair, within a single random subsample, is plotted. The four panels correspond to the risk scores assigned under the post-fairness corrected threshold values, grouped according to protected attribute level, while the three horizontal maroon lines running across the panels correspond to the pre-fairness corrected threshold values.

within the context of the reunification algorithm, the general procedure is broadly applicable both in decision systems using other machine-learning algorithms and in decision systems relying exclusively on human decision-makers. In particular, the first two questions can be asked of any decision-point, whether or not an algorithm informs that decision point. Hence, answering these two questions enables a fairness audit of decision systems driven exclusively by human decision-makers in the same way that it enables a fairness audit of the decisions recommended by a machine learning algorithm. Importantly, however, in the context of a decision system driven exclusively by human decision-makers, the possible answers to the third question are inherently limited to, for example, blunt policy-based procedural changes.

4. Results of fairness correction procedure

The developed procedure discussed in Section 3.3 was applied to the reunification algorithm using 200 random subsamples across each of 101 distinct gated penalty weights (i.e., 0,0.01,…,0.99,1), yielding a set of pre- and post-fairness corrected threshold values at each penalty weight, for each of the three thresholds. To then generate the empirically derived trade-off continuum between fairness and accuracy, the predictive performance and algorithmic fairness of these various sets of threshold values were evaluated.

The dependent nature of the child-reunification pairs again facilitates the need for a subsampling procedure. In particular, for each penalty weight, the pre-fairness corrected threshold values, which are necessarily the same across all penalty weights, and the post-fairness corrected threshold values were used to create risk scores from the predicted probabilities of each of 200 random subsamples of child-reunification pairs. For each such subsample, fairness and predictive performance measures were computed and then averaged across the 200 subsamples. The full details of this procedure are provided in Appendix B.4. For all 101 gated penalty weights, for each of the three thresholds, for both the pre- and post-fairness-corrected threshold values, average Error Rate Balance versus penalty weight is plotted in the left-hand graphic of Fig. 1, while average Error Rate Balance versus the average proportion of changing risk scores (i.e., the empirically derived trade-off continuum between fairness and accuracy) is plotted in the right-hand graphic.

Two observations stand out from Fig. 1. First, fairness is not necessarily maximized when the gated penalty weight is zero, as evidenced by the low- and average-risk thresholds of the left-hand plot. This is not unexpected since the optimization occurs within subsamples, as opposed to across subsamples, and necessitates treating the penalty weight as a tuning parameter, as has been done. Second, the relationship between Error Rate Balance and the proportion of changing risk scores, as evidenced in the right-hand plot, is different and non-linear across all three thresholds. These trade-off continua reveal marginal gains in Error Rate Balance exist beyond 2.5% of risk scores changing, with maximal gains requiring close to 5% of risk scores changing. Such a reality exemplifies

Table 2
The average change (post-fairness corrected minus pre-fairness corrected), rounded to three decimal places, in Error Rate Balance (ERB) and overall predictive performance measures at each threshold, along with corresponding standard deviations across 200 random subsamples. The predictive performance measures include Accuracy (ACC), False Negative Rate (FNR), False Positive Rate (FPR), Negative Predictive Value (NPV), and Positive Predictive Value (PPV).

Measure	Low-Risk Threshold Change		Average-Risk Threshold Change		High-Risk Threshold Change	
	Mean	SD	Mean	SD	Mean	SD
ERB	0.154	0.068	0.218	0.032	0.285	0.034
ACC	−0.012	< 0.0005	−0.006	< 0.0005	0.000	< 0.0005
FNR	−0.005	0.001	−0.011	0.001	−0.004	0.001
FPR	0.015	< 0.0005	0.010	< 0.0005	0.001	< 0.0005
NPV	−0.000	< 0.0005	0.001	< 0.0005	0.001	< 0.0005
PPV	−0.003	< 0.0005	−0.004	0.001	−0.000	0.001

the importance of understanding the functional relationship between fairness and accuracy. In fact, without such knowledge, it is unlikely that stakeholders will confidently identify within the available solution-space the "ideal" trade-off between accuracy and fairness.

4.1. Chosen trade-off

In light of the information conveyed through the plots of Fig. 1, and given that an upper bound of less than 5% of changing risk scores is required to achieve maximum fairness, the "best" post-fairness-corrected threshold values were identified as those that maximize the extent to which Error Rate Balance is achieved. Such maxima were identified at gated penalty weights of 0.17, 0.23, and 0.00 for the low-risk, average-risk, and high-risk thresholds, respectively. Using these values of the gated penalty weight at their respective threshold, the procedure detailed in Appendix B.4 was run one additional time for 200 random subsamples. The resulting average proportion of risk scores that change in transitioning from the corresponding pre- to post-fairness-corrected threshold values is only 4.76% (standard deviation of 0.05%).

To visualize the change in scores resulting from these post-fairness corrected thresholds, consider Fig. 2, where for a single random subsample, each point represents a predicted probability for a unique child-reunification pair. The four panels of this figure correspond to the post-fairness corrected risk scores, grouped according to protected attribute level, and the three horizontal maroon lines correspond to the pre-fairness corrected threshold values. From this plot, the practical impact of the fairness-correction procedure is evident. For example, in panel 3, a small number of BL child-reunification pairs fall above the upper-most maroon line, indicating that these observational units would have been assigned a risk score of 4 under the pre-fairness corrected threshold values, but are now assigned a risk score of 3 under the post-fairness corrected threshold values. Similarly, in that same panel, a small number of HPA child-reunification pairs fall below the middle maroon line, indicating that these observational units would have been assigned a risk score of 2 under the pre-fairness corrected threshold values, but are now assigned a risk score of 3 under the post-fairness corrected threshold values. Furthermore, from this plot it is evident at each threshold that more evidence (i.e., a higher predicted probability) is required of BL and NV child-reunification pairs than of HPA and WH reunification pairs before elevating the corresponding risk score.

4.1.1. Error rate balance and predictive performance assessment

For each threshold, the average change (post-fairness corrected minus pre-fairness corrected) in Error Rate Balance and in five common predictive performance measures as a result of transitioning from the pre- to post-fairness corrected threshold values is provided in Table 2, along with the corresponding standard deviations across the 200 random subsamples. From this table, it is clear that the fairness-correction procedure has, at each threshold, meaningfully improved fairness with comparatively minimal cost to predictive performance. For example, at the high-risk threshold, the minimum parity in both the false positive and false negative error rates between any two levels of the protected attribute (i.e., ERB) has been increased, on average, by 0.285 (standard deviation of 0.034), with the cost of this improvement, on average, being no more than 0.004, in absolute value, to any of the overall predictive performance measures at that threshold. Importantly, the fairness at each threshold is not only improved, but categorically good with an average Error Rate Balance of 0.83, 0.82, and 0.87 at the low-, average-, and high-risk thresholds, respectively.

4.2. Considering alternative group-level definitions

For jurisdictions that elect, or use-cases that result in, an alternative definition of algorithmic fairness, the presented procedure is applicable. In Appendix C, we demonstrate such generalizability within the context of the reunification algorithm, applying the procedure under three alternative specifications of algorithmic fairness: Conditional Use Accuracy Equality, Treatment Equality, and Calibration. Regardless of the utilized definition, however, difficult trade-offs will persist. In fact, these difficult trade-offs exist even without the use of an algorithm, and "[r]ejecting model-driven or automated decision making is not a way to avoid these problems" (Mitchell, Potash, Barocas, & Alexander D'Amour, 2020, pg. 15).

Hence, rather than accepting the "default" trade-offs of either purely human-based decision systems or uncorrected classification algorithms, we advocate for identifying the trade-offs that are most appropriate for the use-case, prior to training the algorithm, and then fairness-correcting accordingly. We hope that the broad generalizability of the presented procedure can help move conversations surrounding the algorithmic fairness of Child Welfare machine learning tools from broadly stated concerns to direct inquires pertaining to the choice of the protected attributes, the definition of algorithmic fairness, and the utilized correction procedure.

5. Discussion

A first-of-its-kind machine learning algorithm designed to serve as a decision-support tool for Oregon Child Welfare staff tasked with making permanency-related decisions was developed and presented. This algorithm estimates the probability of a child re-entering substitute care within one year of an initiated reunification. This outputted probability is then thresholded to provide front-line staff a corresponding four-tier risk score. Importantly, using the procedure developed and presented above, which is a contextually necessary extension of procedures proposed in the literature, these risk scores have been "corrected" to mitigate a lack of algorithmic fairness across a race and ethnicity-based protected attribute. The presented reunification tool is therefore also the first Child Welfare machine learning algorithm to proactively address algorithmic fairness. We hope, broadly speaking, this work will subsequently help set a new standard of practice for algorithmic fairness in the machine learning tools of Child Welfare.

While the developed tool represents a substantial step forward from the otherwise current "standards" of algorithmic fairness in the predictive analytic tools of Child Welfare, it is far from a proverbial "final" step. We highlight below some additional efforts that could lead to further progress. Note that these highlighted efforts, like the presented work above, assume that the decision to develop an algorithm has been made, with an identified outcome and feature set, and will therefore require decisions surrounding the protected attribute(s), definition of algorithmic fairness, and corresponding fairness-correction procedure. This does not mean, however, that there are not important choices and assumptions preceding this stage, including the choice of outcome variable, which have ramifications for the algorithmic fairness of the final developed tool or whether such a tool is pursued in the first place (e.g.,

Green (2020), Mitchell et al. (2020), Obermeyer, Powers, Vogeli, & Mullainathan (2019)). Such important considerations, however, are beyond the focus of this current paper.

5.1. Future exploration

5.1.1. Cross-category definitions of algorithmic fairness

The choice of a group-level definition in this work was driven by the current limitations of individual-level and causal-reasoning based definitions of algorithmic fairness. However, there is potential for blending the ideas and concepts from all three of these categories into an algorithmic fairness approach. For example, in Chouldechova and Roth (2018), several proposed approaches seeking to bridge the gap between group- and individual-level fairness are highlighted, with source references provided, where the objective is to identify statistical fairness definitions which hold "…not just on a small number of protected groups, but on an exponential or infinite class of groups" (Chouldechova & Roth, 2018, pg.4). Whether such approaches will ultimately be viable is yet unknown, but if successful these approaches could yield the benefits of both categories of definitions.

Furthermore, while notions of causal-reasoning based definitions of fairness are currently not a pragmatically viable option, the exercise of considering fairness through such a frame "…can make value judgments more explicit…[and] allows practitioners to designate which causal pathways from sensitive attributes to decisions constitute 'acceptable' or 'unacceptable' sources of dependence between sensitive attributes and decisions" (Mitchell et al., 2020, pg.13). In other words, the framework provided through such approaches, even if not directly implementable, can still serve (1) to potentially inform which features are included in the training set and (2) to potentially filter which group-level definitions are viable within a given use-case. While this does not bypass the difficulties of attempting to map out the hypothetical causal paths in proxy-laden administrative data, it does suggest that embarking on such an effort may have utility.

5.1.2. Multi-stage algorithmic fairness correction procedures

The use of the single "correction" procedure developed and presented in this paper was motivated by the self-imposed transparency, interpretability, and operationalizability characteristics discussed in Section 3.3. Such motivations precluded the exploration of pre- and in-processing approaches, which may yield "better" trade-offs between algorithmic fairness and predictive performance, albeit at the expense of some combination of these three characteristics. We suspect that in the future, however, when algorithmic fairness considerations have become innate to algorithm development within Child Welfare, the need for such characteristics may be softened. In such a future, because it is possible within a single use-case to combine approaches across pre-, in-, and post-processing classes (Berk et al., 2017), the search for "best" solutions could then include the exploration of composite correction procedure, which is analogous to the common machine learning techniques of model ensembling and model stacking. This could lead to a multi-staged algorithmic fairness correction procedure in which, for example, a pre-processing procedure is first applied before training the algorithm and a post-processing procedure is also applied at the thresholding stage.

5.1.3. Fairness across sequential decisions

While the reunification algorithm developed and presented in this paper is viewed through the lens of a single decision point, there are critical decision points within Child Welfare that come before and, potentially, after this decision point. If other machine learning algorithms are developed to support human decision-makers at these other decision points, and assuming these tools are also "corrected" to mitigate a lack of algorithmic fairness in the generated risk scores, what considerations should be given to the algorithmic fairness across the composition of these decision points? The need to explore algorithmic fairness across a sequence of decision points, rather than within an isolated decision point, is highlighted in Chouldechova and Roth (2018) and initially explored in Bower et al. (2017) under the label of fair pipelines. Such work will likely garner greater attention after more jurisdictions have developed and deployed a single machine learning algorithm and correspondingly started contemplating additional algorithms.

5.2. Final thoughts

The reunification tool and corresponding algorithmic fairness correction procedure presented in this work are far from perfect, but together they can help improve upon the status quo by equitably increasing the amount and success rate of reunifications. While the use of algorithms in domains such as Child Welfare is hotly contested, it should be recognized that in use-cases such as this, human decision-makers access and are informed by the same administrative data that the algorithm is trained on. Furthermore, the extent to which that administrative data influences their decisions is understandably and unavoidably inconsistent across observational units. Hence, by implementing a risk algorithm, like the reunification tool, to inform human-decisions, this inconsistent and incomplete use of administrative data by staff is buttressed with a decision support tool that not only leverages administrative data in a consistent and thorough manner, but also adheres to an agreed upon set of shared values articulated through the identified definition of algorithmic fairness.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Modeling best practices

A.1. Selective labeling

The selective labeling problem (Lakkaraju, Kleinberg, Leskovec, Ludwig, and Mullainathan, 2017) occurs when historical data are overly influenced by the very decision which a tool seeks to support. In Child Welfare reunification, this occurs because outcomes (e.g., whether the child experiences further abuse/neglect at home) is partly influenced by the very act of returning the child home. Thus, the historical data should be analyzed solely for child-transition pairs which represent a reunification with family. Thus, the outcomes of interest become conditional on the decision, and the machine learning classifier is less likely to perpetuate poor decisions. In this way, when we refer to a child welfare outcome, we are sure to avoid bias via the selective label problem by ensuring that we are calculating the likelihood of the outcome conditional on the decision (e.g., the likelihood of a return to substitute care conditional on the child's reunification with his/her family).

A.2. Repeated observations

Repeated observation leakage can occur if independence is violated in the data used to train the machine learning classifier. In Child Welfare reunification, children may have been involved in multiple reunification events over time. Thus, it is important to unduplicate the historical data so that the same child does not occur twice in the data set. When the data is split into training data and testing data, in order to test the predictive performance of the machine learning classifier, this unduplication ensures that information does not "leak" between the training set and the testing set, which can cause predictive performance to be inflated. Because the unduplication reduces the volume of the historical data, we have constructed a repeated sample technique which repeats the unduplication step over and over until a series of classification models are constructed and all child-reunification pairs have been included in at least one of the models. This technique is rooted in the common data science practice of model ensembling. It ensures that we can maximally draw on historical data in the construction of our decision support tool, and do so in a way that prevents the inflation of predictive performance metrics.

A.3. Censoring and data windows

Oregon's Child Welfare's SACWIS system came online in August 2011. This means that historical data is available in a reliable and consistent format only from a certain historical time point. Reunifications that occurred in February 2012 have only six months of reliable historical data, while reunifications that occurred in August 2019 have eight years of reliable historical data. For this reason, it is important to standardize the historical data window. Otherwise, the date of a historical reunification will have undue influence on the historical calculation of risk, and the model will not be generalizable to new observations. To prevent this from occurring, we set a historical data window to 1.5 years.

A similar censoring bias can occur with the outcome window. For example, a child involved in a February 2012 reunification has eight years of time in which to experience (or not experience) an outcome, whereas a child involved in an August 2019 reunification has had only one year to experience (or not experience) the outcome. Consequently, a fixed outcome window of one year was established to prevent undue data bias via the timing of the report.

Appendix B. Full methodological details

The primary objective of this section is to provide the full methodological details of the developed algorithmic fairness correction procedure. To help facilitate understanding of the details of the procedure, and in particular how it generalizes to eight other group-level definitions of algorithmic fairness, a rigorous exemplification of the approach to quantifying such fairness measures is first required.

B.1. Quantifying error rate balance

To quantify the extent to which Error Rate Balance is achieved for a single random subsample of child-reunification pairs, we identify at each threshold the greatest disparity in either false negative rates or false positive rates across any pairing of protected attribute levels. To help illustrate this measure, consider Table 3. The left-hand side of this table provides, for each level of the protected attribute, the false positive and false negative rates at the average risk threshold for a single random subsample of child-reunification pairs. Correspondingly, the right-hand side of this table provides, for all possible pairings of protected attribute levels, the pairwise ratio of false negative rates and the pairwise ratio of false positive rates.

To quantify the extent of the observed disproportionality in these false negative and false positive rates, we utilize pairwise error rate ratios. Consider, for example, the pairwise false positive rate ratio between BL and HPA child-reunification pairs that is provided in row 2, column 1 of the right-hand matrix of Table 3. The given value of 0.60 is obtained by taking the smaller false positive rate between the BL and HPA levels and dividing it by the larger of the two false positive rates, i.e., $\frac{0.203}{0.342} \approx 0.60$. This value conveys that the rate at which would-be-successful HPA child-reunification pairs are incorrectly predicted to "fail" is only 0.60 the rate at which would-be-successful BL child-reunification pairs are incorrectly predicted to "fail". As yet another example, consider the pairwise false negative rate ratio between NV and WH child-reunification pairs that is provided in row 3, column 4 of the right-hand matrix of Table 3. The given value of 0.80 is obtained by taking the smaller false negative rate between the NV and WH levels and dividing it by the larger of the two false negative rates, i.e., $\frac{0.309}{0.386} \approx 0.80$. This value conveys that the rate at which would-be-unsuccessful NV child-reunification pairs are incorrectly predicted to "succeed" is only 0.80 the rate at which would-be-unsuccessful WH child-reunification pairs are incorrectly predicted to "succeed". These two values, along with the other 10 analogously calculated ratios, each individually shed light on the potential unfairness of the algorithm at the average-risk threshold, but all 12 collectively must be considered in performing a comprehensive assessment.

To achieve such an assessment, we summarize these 12 pairwise error rate ratios into a single number. In choosing this summary measure,

Table 3
For each level of the protected attribute, for a single random subsample of child-reunification pairs, the false positive and false negative rates at the average risk threshold are provided in the left-hand side of this table. Correspondingly, in the right-hand side, the pairwise ratio of false negative rates and the pairwise ratio of false positive rates are provided for all possible pairings of protected attribute levels. Note that these pairwise error rate ratios are always constructed such that the larger error rate is in the denominator and the smaller is in the numerator, thus ensuring that all ratios are between 0 and 1. To then quantify with a single number the extent to which Error Rate Balance is achieved for a single random subsample at an arbitrary threshold, we utilize the minimum of the 12 pairwise error rate ratios, which corresponds to 0.60 in the table below.

Level	*Error Rate*			*Pairwise Error Rate Ratios*			
	FNR	FPR		BL	HPA	NV	WH
BL	0.331	0.342	BL		0.90	0.93	0.86
HPA	0.368	0.203	HPA	0.60		0.84	0.95
NV	0.309	0.310	NV	0.91	0.66		0.80
WH	0.386	0.279	WH	0.82	0.73	0.90	

recognize that all pairwise error rate ratios will always be between 0 and 1, where a value of 1 means that the two corresponding levels of the protected attribute have the same false negative rate, or the same false positive rate, depending on which error rate is being considered. Furthermore, the smaller the pairwise error rate ratio, the greater the disproportionality in the respective error rates between the two corresponding levels of the protected attribute. The smallest of these 12 ratios therefore represents the most egregious instance of disproportionality in error rates across the levels of the protected attribute. Consequently, to quantify with a single number the extent to which Error Rate Balance is achieved, for a single random subsample, at any specified threshold, we utilize the minimum of the twelve possible pairwise error rate ratios; for the subsample yielding Table 3, this value is 0.60. The developed procedure to mitigate a lack of Error Rate Balance seeks to increase this measure towards one. To accommodate the dependence structure of the observational units, this Error Rate Balance measure is then calculated, at each threshold, across a large number of random subsamples. The corresponding average value for each threshold represents an audit of the extent to which Error Rate Balance is achieved at that threshold.

B.1.1. Quantifying other group-level measures of fairness

The approach to quantifying the extent to which Error Rate Balance is achieved across a single subsample of child-reunification pairs is straightforwardly generalized for utilization with seven other group-level definitions of fairness. In particular, this quantification approach can be directly utilized with Statistical Parity, Overall Accuracy Equality, Predictive Parity, Equal Opportunity, Predictive Equality, Conditional Use Accuracy Equality, and Treatment Equality by replacing the use of pairwise ratios across false negative and false positive rates with the corresponding pairwise ratios across analogous measure(s) dictated by the particular definition of algorithmic fairness. For example, with Conditional Use Accuracy Equality, these measures would be the positive predictive and negative predictive values, whereas with Predictive Parity this measure would be just the positive predictive value.

With one alternative group-level definition, Calibration, the generalization is subtly, but importantly, different. In particular, with Calibration, the requisite pairwise ratios are computed at each risk score rather than at each threshold. Consequently, the approach amounts to first calculating the proportion of observational units assigned risk score S, where $S \in \{1, 2, 3, 4\}$, that go on to experience the adverse event, and then calculating the pairwise ratios of such a proportion across each pairing of protected attribute levels..

B.2. Full fairness correction procedure

In this section, we provide the step-by-step details of the proposed process for obtaining post-fairness-corrected threshold values within the context of the reunification algorithm. While the procedure is exemplified for three thresholds – low-risk, average-risk, and high-risk – and four protected attribute levels – BL, HPA, NV, and WH – the procedure can straightforwardly be generalized to accommodate an arbitrary number of thresholds and protected attribute levels.

Let X be the $4 \times N$ matrix in which the N rows represent the entire sample of child-reunification pairs and the 4 columns represent, respectively, a unique child identification number (for subsampling purposes), the observed value of the binary outcome variable, the level of the protected attribute, and the predicted probability obtained from the algorithm.

1. Set number of subsamples, I (e.g., $I = 200$).
2. Set the value of the tuning parameter, w, where $0 \leqslant w \leq 1$.
3. Initialize the subsample index: $i = 1$.
4. Obtain the i^{th} random subsample of $\boldsymbol{X}$, denoted by $\boldsymbol{X}_i$.
5. Calculate the group-agnostic low-risk, average-risk, and high-risk threshold values, $\Phi_{L,i}, \Phi_{A,i}, \Phi_{H,i}$, corresponding to subsample i, where
   - $\Phi_{A,i}$ represents the average predicted probability for the child-reunification pairs of subsample i,
   - $\Phi_{L,i}$ represents the 50^{th} percentile of predicted probabilities for the child-reunification pairs of subsample i that are less than $\Phi_{A,i}$, and
   - $\Phi_{H,i}$ represents the 75^{th} percentile of predicted probabilities for the child-reunification pairs of subsample i that are greater than $\Phi_{A,i}$.

   † Recall that the identification of such cutoffs is rooted in the use-case.
6. Obtain the group-specific low-risk threshold values for subsample $i, \widehat{\boldsymbol{\theta}_{L,i}} = \left(\widehat{\theta}_{L,i}^{BL}, \widehat{\theta}_{L,i}^{HPA}, \widehat{\theta}_{L,i}^{NV}, \widehat{\theta}_{L,i}^{WH}\right)'$, by solving the following penalized optimization problem

$$\begin{aligned} \underset{\boldsymbol{\theta}_{L,i}}{\text{argmin}} \quad & \left(1 - w\right)\left(1 - \text{ERB}\left(\boldsymbol{\theta}_{L,i}\right)\right) \; + \; w\Delta\left(\boldsymbol{\theta}_{L,i}, \Phi_{L,i}\right) \\ \text{subject to} \quad & 0 < \min\left\{\theta_{L,i}^{BL}, \theta_{L,i}^{HPA}, \theta_{L,i}^{NV}, \theta_{L,i}^{WH}\right\} \leqslant \Phi_{L,i} \\ & \Phi_{L,i} \leqslant \max\left\{\theta_{L,i}^{BL}, \theta_{L,i}^{HPA}, \theta_{L,i}^{NV}, \theta_{L,i}^{WH}\right\} < \Phi_{A,i}, \end{aligned} \tag{1}$$

   where $\text{ERB}(\boldsymbol{\theta}_{L,i})$ is the value quantifying the extent to which Error Rate Balance is achieved at $\boldsymbol{\theta}_{L,i}$, and $\Delta(\boldsymbol{\theta}_{L,i}, \Phi_{L,i})$ is the proportion of risk scores that change (i.e., either from 1 to 2 or from 2 to 1) when moving from the group-agnostic threshold value, $\Phi_{L,i}$, to the group-specific threshold values specified by $\boldsymbol{\theta}_{L,i}$.
7. Obtain the group-specific average-risk threshold values for subsample $i, \widehat{\boldsymbol{\theta}_{A,i}} = \left(\widehat{\theta}_{A,i}^{BL}, \widehat{\theta}_{A,i}^{HPA}, \widehat{\theta}_{A,i}^{NV}, \widehat{\theta}_{A,i}^{WH}\right)'$, by solving the following penalized optimization problem

$$
\begin{aligned}
\underset{\boldsymbol{\theta}_{A,i}}{\text{argmin}} \quad & (1-w)\left(1-\text{ERB}\left(\boldsymbol{\theta}_{A,i}\right)\right) \; + \; w\Delta\left(\boldsymbol{\theta}_{A,i}, \Phi_{A,i}\right) \\
\text{subject to} \quad & \widehat{\theta}_{L,i}^{BL} < \quad \theta_{A,i}^{BL} \leqslant \Phi_{H,i} \\
& \widehat{\theta}_{L,i}^{HPA} < \quad \theta_{A,i}^{HPA} \leqslant \Phi_{H,i} \\
& \widehat{\theta}_{L,i}^{NV} < \quad \theta_{A,i}^{NV} \leqslant \Phi_{H,i} \\
& \widehat{\theta}_{L,i}^{WH} < \quad \theta_{A,i}^{WH} \leqslant \Phi_{H,i} \\
& \min\left\{\theta_{A,i}^{BL}, \theta_{A,i}^{HPA}, \theta_{A,i}^{NV}, \theta_{A,i}^{WH}\right\} \leqslant \Phi_{A,i} \\
& \max\left\{\theta_{A,i}^{BL}, \theta_{A,i}^{HPA}, \theta_{A,i}^{NV}, \theta_{A,i}^{WH}\right\} \geqslant \Phi_{A,i},
\end{aligned}
\tag{2}
$$

where $\text{ERB}(\boldsymbol{\theta}_{A,i})$ is the value quantifying the extent to which Error Rate Balance is achieved at $\boldsymbol{\theta}_{A,i}$, and $\Delta(\boldsymbol{\theta}_{A,i}, \Phi_{A,i})$ is the proportion of risk scores that change (i.e., either from 2 to 3 or from 3 to 2) when moving from the group-agnostic threshold value, $\Phi_{A,i}$, to the group-specific threshold values specified by $\boldsymbol{\theta}_{A,i}$.

8. Obtain the group-specific high-risk threshold values for subsample i, $\widehat{\boldsymbol{\theta}_{H,i}} = \left(\widehat{\theta}_{H,i}^{BL}, \widehat{\theta}_{H,i}^{HPA}, \widehat{\theta}_{H,i}^{NV}, \widehat{\theta}_{H,i}^{WH}\right)'$, by solving the following penalized optimization problem

$$
\begin{aligned}
\underset{\boldsymbol{\theta}_{A,i}}{\text{argmin}} \quad & (1-w)\left(1-\text{ERB}\left(\boldsymbol{\theta}_{H,i}\right)\right) \; + \; w\Delta\left(\boldsymbol{\theta}_{H,i}, \Phi_{H,i}\right) \\
\text{subject to} \quad & \widehat{\theta}_{A,i}^{BL} < \quad \theta_{H,i}^{BL} < 1 \\
& \widehat{\theta}_{A,i}^{HPA} < \quad \theta_{H,i}^{HPA} < 1 \\
& \widehat{\theta}_{A,i}^{NV} < \quad \theta_{H,i}^{NV} < 1 \\
& \widehat{\theta}_{A,i}^{WH} < \quad \theta_{H,i}^{WH} < 1 \\
& \min\left\{\theta_{H,i}^{BL}, \theta_{H,i}^{HPA}, \theta_{H,i}^{NV}, \theta_{H,i}^{WH}\right\} \leqslant \Phi_{H,i} \\
& \max\left\{\theta_{H,i}^{BL}, \theta_{H,i}^{HPA}, \theta_{H,i}^{NV}, \theta_{H,i}^{WH}\right\} \geqslant \Phi_{H,i},
\end{aligned}
\tag{3}
$$

where $\text{ERB}(\boldsymbol{\theta}_{H,i})$ is the value quantifying the extent to which Error Rate Balance is achieved at $\boldsymbol{\theta}_{H,i}$, and $\Delta(\boldsymbol{\theta}_{H,i}, \Phi_{H,i})$ is the proportion of risk scores that change (i.e., either from 3 to 4 or from 4 to 3) when moving from the group-agnostic threshold value, $\Phi_{H,i}$, to the group-specific threshold values specified by $\boldsymbol{\theta}_{H,i}$.

9. Increase subsample index: $i = i + 1$. If $i \leqslant I$, repeat steps 4–8, else move on to step 10.
10. For the specified value of w, obtain the pre-fairness corrected values for the low-risk threshold, $\Phi_{L,w}$, the average-risk threshold, $\Phi_{A,w}$, and the high-risk threshold, $\Phi_{H,w}$. These values are obtained via the following bagging-like process:
    - $\Phi_{L,w} = \frac{1}{I}\sum_{i=1}^{I}\Phi_{L,i}$, with $\Phi_{A,w}$ and $\Phi_{H,w}$ analogously calculated.
11. For the specified value of w, obtain the post-fairness corrected values for the group-specific low-risk thresholds, $\widehat{\boldsymbol{\theta}_{L,w}} = \left(\widehat{\theta}_{L,w}^{BL}, \widehat{\theta}_{L,w}^{HPA}, \widehat{\theta}_{L,w}^{NV}, \widehat{\theta}_{L,w}^{WH}\right)'$, the group-specific average-risk thresholds, $\widehat{\boldsymbol{\theta}_{A,w}} = \left(\widehat{\theta}_{A,w}^{BL}, \widehat{\theta}_{A,w}^{HPA}, \widehat{\theta}_{A,w}^{NV}, \widehat{\theta}_{A,w}^{WH}\right)'$, and the group-specific high-risk thresholds, $\widehat{\boldsymbol{\theta}_{H,w}} = \left(\widehat{\theta}_{H,w}^{BL}, \widehat{\theta}_{H,w}^{HPA}, \widehat{\theta}_{H,w}^{NV}, \widehat{\theta}_{H,w}^{WH}\right)'$. These values are obtained via the following bagging-like process:
    - $\widehat{\theta}_{L,w}^{BL} = \frac{1}{I}\sum_{i=1}^{I}\widehat{\theta}_{L,i}^{BL}$, with $\widehat{\theta}_{L,w}^{HPA}$, $\widehat{\theta}_{L,w}^{NV}$, and $\widehat{\theta}_{L,w}^{WH}$ similarly calculated. The constituent components of $\widehat{\boldsymbol{\theta}_{A,w}}$ and $\widehat{\boldsymbol{\theta}_{H,w}}$ are analogously calculated as well.

B.2.1. Understanding the constraints of the optimization procedure

This section clarifies the objectives of the constraints associated with the optimization problem in steps 6–8 of Appendix B.2. In particular, through these constraints the identified group-specific threshold values are guaranteed to exist (i.e., be between 0 and 1) and to exhibit two properties we label as "orderliness" and "coveredness". By orderliness, we mean that within a particular level of the protected attribute, it must be the case that the low-risk threshold value is less than the average-risk threshold value, which is in turn less than the high-risk threshold value (e.g., $\widehat{\theta}_{L,i}^{BL} < \widehat{\theta}_{A,i}^{BL} < \widehat{\theta}_{H,i}^{BL}$). By coveredness, we mean that at a particular threshold, the group-agnostic threshold value must be within the closed interval created by the corresponding minimum and maximum group-specific threshold values (e.g., $\Phi_{A,i} \in [\min\widehat{\boldsymbol{\theta}_{A,i}}, \max\widehat{\boldsymbol{\theta}_{A,i}}]$). The logic with this coveredness property is to retain, as much as possible, the meaning and intent behind the "original", group-agnostic, threshold values that were identified as part of the business case motivating the algorithm's development.

B.2.2. Understanding the procedure

This section provides insight into the logic of the correction procedure. The subsampling component of this procedure addresses the dependence structure of the observational units (i.e., child-reunification pairs). in instances where no such dependence structure exists, two options are possible. In particular, either the procedure could be run one time (i.e., $I = 1$) on the full data set, or a bootstrap resampling approach could implemented with I equal to the number of desired bootstrap resamples.

To recognize how the procedure yields an empirical curve of the trade-off continuum between accuracy and fairness, consider the penalized objective function utilized with each subsample. This function measures the extent of the algorithmic unfairness in subsample i through the

$1 - ERB(\boldsymbol{\theta}_{t,i})$ term. Since smaller values of this term correspond to "fairer" group-specific threshold values, the "fairest" threshold values are obtained by minimizing this term. To then quantify and capture the trade-off between fairness and accuracy, this term is penalized by the corresponding decrease in accuracy incurred as a result of adopting the group-specific threshold values over the group-agnostic threshold value. The utilized penalty term, $\Delta(\boldsymbol{\theta}_{t,i}, \Phi_{t,i})$, is not a direct measure of any specific type of predictive accuracy, but is instead a broad catch-all for all such measures. In particular, when a risk score changes as a result of transitioning from the group-agnostic threshold value to the group-specific threshold values, all predictive performance measures are necessarily impacted, though not necessarily for the worse. Hence, the greater the proportion of changing risk scores, the greater the potential accuracy-related costs.

The gated weighting mechanism, with weight $0 \leqslant w \leq 1$, bounds how far from the group-agnostic threshold value the optimizer is willing to search for the optimal group-specific threshold values. When $w = 0$, the optimization process will yield the fairest group-specific threshold values for subsample i without any regard for accuracy. When $w = 1$, the optimization process in exclusively concerned with accuracy and, consequently, the returned group-specific threshold values for subsample i will be identical to the group-agnostic threshold value. More generally, as the value of w is increased over the range of values between these extremes, the pursuit of fairness becomes more heavily anchored to the accuracy achieved through the group-agnostic threshold value. By iterating the procedure across a tuning grid of w-values ranging from zero to one, the functional relationship between accuracy and fairness can empirically be explored at each threshold.

Finally, the procedure is easily applied across alternative group-level definitions of algorithmic fairness. This is achieved by replacing $ERB(\boldsymbol{\theta}_{t,i})$ with the analogous measure for the specified alternative definition, as discussed in Appendix B.1.1.

B.2.3. A visual check on the procedure

This section provides a visual means for better understanding key components of the procedure. In particular, recall that the procedure described in Section B.2 was applied to the reunification algorithm with 200 random subsamples (i.e., $I = 200$) utilized across each of 101 distinct penalty weights (i.e., $w = 0, 0.01, \ldots, 0.99, 1$). Fig. 3 displays the corresponding pre- and post-fairness-corrected low-risk threshold values obtained at each penalty weight; analogous results are found at both the average-risk and high-risk thresholds. Observe in this figure that, as should necessarily be the case, the pre-fairness-corrected threshold values are the same across all four protected attribute levels and all penalty weights. Furthermore, observe that as w increases towards one, the distance between the pre- and post-fairness-corrected threshold values tends to decrease towards zero. This is by design and demonstrates that the penalty term in the penalized objective function is operating as intended. More specifically, as costs to accuracy become more and more "valued", as conveyed through the increasing value of w, the post-fairness-corrected threshold values are "pulled back" towards their corresponding "accuracy-anchored" pre-fairness-corrected threshold value. Finally, for each value of w, observe the intended consequence of the "coveredness" constraint in that at least one protected attribute level has their post-fairness corrected threshold value above the pre-fairness corrected threshold value and at least one protected attribute levels has their post-fairness corrected threshold value below the pre-fairness corrected threshold value.

B.3. Simultaneous penalized optimization for calibration

The step-by-step procedure detailed in Appendix B.2 directly applies to eight of the nine group-level definitions of algorithmic fairness discussed in this manuscript. More specifically, in addition to Error Rate Balance, the procedure can be directly utilized with Statistical Parity, Overall Accuracy Equality, Predictive Parity, Equal Opportunity, Predictive Equality, Conditional Use Accuracy Equality, and Treatment Equality by replacing $\text{ERB}(\boldsymbol{\theta}_{\cdot,i})$ with an analogous measure dictated by the particular definition of algorithmic fairness, as discussed in Appendix B.1.1.

For Calibration, however, the sequential penalized optimization problems in steps 6–8 of Appendix B.2 must be combined into one simultaneous penalized optimization problem, given below. Additionally, note that $\text{ERB}(\boldsymbol{\theta}_{\cdot,i})$, which measures the extent two which Error Rate Balance is achieved at an arbitrary threshold for subsample i (so $\boldsymbol{\theta}_{\cdot,i} = \boldsymbol{\theta}_{L,i}$, $\boldsymbol{\theta}_{A,i}$, or $\boldsymbol{\theta}_{H,i}$), is now replaced with $\text{CAL}(S_i, \boldsymbol{\theta}_{T,i})$, which measures the extent to which Calibration is achieved at risk score $S \in \{S1, S2, S3, S4\}$ for subsample i, which depends on one or more thresholds in the set of all thresholds (so $\boldsymbol{\theta}_{T,i} = (\boldsymbol{\theta}_{L,i}, \boldsymbol{\theta}_{A,i}, \boldsymbol{\theta}_{H,i})'$). Calculation of $\text{CAL}(S_i, \boldsymbol{\theta}_{T,i})$ at each risk score $S \in \{S1, S2, S3, S4\}$ is described in Appendix B.1.1. Consequently, Calibration can be utilized within the developed fairness correction procedure by replacing steps 6–8 of Appendix B.2 with the following step:

- Obtain the group-specific threshold values across the low-, average-, and high-risk thresholds, for the i^{th} subsample, $\widehat{\boldsymbol{\theta}_{T,i}} = \left(\widehat{\boldsymbol{\theta}_{L,i}}, \widehat{\boldsymbol{\theta}_{A,i}}, \widehat{\boldsymbol{\theta}_{H,i}}\right)'$, where
  $\widehat{\boldsymbol{\theta}_{L,i}} = \left(\widehat{\theta}^{BL}_{L,i}, \widehat{\theta}^{HPA}_{L,i}, \widehat{\theta}^{NV}_{L,i}, \widehat{\theta}^{WH}_{L,i}\right)'$,
  $\widehat{\boldsymbol{\theta}_{A,i}} = \left(\widehat{\theta}^{BL}_{A,i}, \widehat{\theta}^{HPA}_{A,i}, \widehat{\theta}^{NV}_{A,i}, \widehat{\theta}^{WH}_{A,i}\right)'$, and.
  $\widehat{\boldsymbol{\theta}_{H,i}} = \left(\widehat{\theta}^{BL}_{H,i}, \widehat{\theta}^{HPA}_{H,i}, \widehat{\theta}^{NV}_{H,i}, \widehat{\theta}^{WH}_{H,i}\right)'$, by solving the following penalized optimization problem

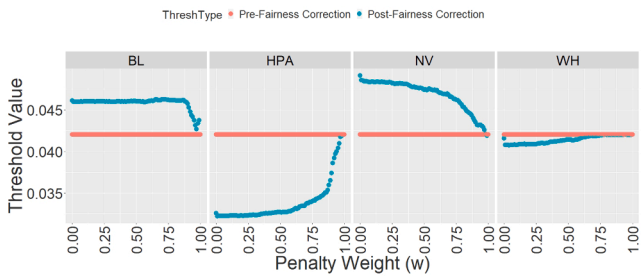


Fig. 3. The pre- and post-fairness corrected low-risk threshold values resulting from utilizing the procedure detailed in Section B.2 with 200 random subsamples (i.e., $I = 200$) across each of 101 distinct penalty weights (i.e., $w = 0, 0.01, 0.02, \ldots, 0.98, 0.99, 1$).

$$
\begin{aligned}
\underset{\theta_{T,i}}{\operatorname{argmin}} \quad & w\Delta\left(\boldsymbol{\theta}_{T,i}, \boldsymbol{\Phi}_{T,i}\right) + \left(1-w\right)\sum_{S\in\{S1,S2,S3,S4\}}\left(1-\mathrm{CAL}\left(S_i, \boldsymbol{\theta}_{T,i}\right)\right) \\
\min\left\{\theta_{L,i}^{BL}, \theta_{L,i}^{HPA}, \theta_{L,i}^{NV}, \theta_{L,i}^{WH}\right\} \leqslant & \Phi_{L,i} \\
\text{subject to} \quad & \max\left\{\theta_{L,i}^{BL}, \theta_{L,i}^{HPA}, \theta_{L,i}^{NV}, \theta_{L,i}^{WH}\right\} \geqslant \Phi_{L,i} \\
& \min\left\{\theta_{A,i}^{BL}, \theta_{A,i}^{HPA}, \theta_{A,i}^{NV}, \theta_{A,i}^{WH}\right\} \leqslant \Phi_{A,i} \\
& \max\left\{\theta_{A,i}^{BL}, \theta_{A,i}^{HPA}, \theta_{A,i}^{NV}, \theta_{A,i}^{WH}\right\} \geqslant \Phi_{A,i} \\
& \min\left\{\theta_{H,i}^{BL}, \theta_{H,i}^{HPA}, \theta_{H,i}^{NV}, \theta_{H,i}^{WH}\right\} \leqslant \Phi_{H,i} \\
& \max\left\{\theta_{H,i}^{BL}, \theta_{H,i}^{HPA}, \theta_{H,i}^{NV}, \theta_{H,i}^{WH}\right\} \geqslant \Phi_{H,i} \\
& 0 < \quad \theta_{L,i}^{BL} \quad < \quad \theta_{A,i}^{BL} \quad < \quad \theta_{H,i}^{BL} < 1 \\
& 0 < \quad \theta_{L,i}^{HPA} \quad < \quad \theta_{A,i}^{HPA} \quad < \quad \theta_{H,i}^{HPA} < 1 \\
& 0 < \quad \theta_{L,i}^{NV} \quad < \quad \theta_{A,i}^{NV} \quad < \quad \theta_{H,i}^{NV} < 1 \\
& 0 < \quad \theta_{L,i}^{WH} \quad < \quad \theta_{A,i}^{WH} \quad < \quad \theta_{H,i}^{WH} < 1,
\end{aligned}
\tag{4}
$$

where $\boldsymbol{\Phi}_{T,i} = (\Phi_{L,i},\ \Phi_{A,i}\ \Phi_{H,i})'$ and $\Delta(\boldsymbol{\theta}_{T,i}, \boldsymbol{\Phi}_{T,i})$ is the proportion of risk scores that change in transitioning from the group-agnostic to group-specific threshold values.

B.4. Generating empirically derived trade-off continuum

Lastly, in this section we present the step-by-step details for generating various fairness and predictive performance measures necessary for generating, among other things, the empirically derived trade-off continuum between accuracy and fairness. Ultimately, this amounts to iterating through each value of w and completing the following steps.

1. Set number of subsamples, J (e.g., $J = 200$).
2. Initialize the subsample index: $j = 1$.
3. Obtain the j^{th} random subsample of $\boldsymbol{X}$, denoted by $\boldsymbol{X}_j$.
4. Calculate the extent to which Error Rate Balance is achieved for the j^{th} subsample at the pre-fairness corrected low-risk threshold value, denoted $\mathrm{ERB}(\Phi_{L,w}, \boldsymbol{X}_j)$, and at the post-fairness corrected low-risk threshold values, denoted $\mathrm{ERB}(\widehat{\boldsymbol{\theta}_{L,w}}, \boldsymbol{X}_j)$. Additionally, perform the analogous calculations at the average-risk and high-risk thresholds.
5. Calculate various performance measures for the j^{th} subsample at the pre-fairness corrected low-risk threshold value, $\Phi_{L,w}$, and at the post-fairness corrected low-risk threshold values, $\widehat{\boldsymbol{\theta}_{L,w}}$. These performance measures include the false negative rate, true positive rate, false positive rate, true negative rate, positive predictive value, false discovery rate, negative predictive value, false omission rate, and accuracy for each level of the protected attribute, as well as overall. Additionally, perform the analogous calculations at the average-risk and high-risk thresholds.
6. Increase subsample index: $j = j + 1$. If $j \leqslant J$, repeat steps 3–5, else move on to step 7.
7. Obtain the average and standard deviation of extent to which Error Rate Balance is achieved across the J subsamples at both the pre-fairness corrected value for the low-risk threshold and the post-fairness corrected values for the low-risk threshold. Obtain analogous measures for the average-risk and high-risk thresholds. These values are calculated as follows:
   - $\overline{\mathrm{ERB}}\left(\Phi_{L,w}, \boldsymbol{X}\right) = \frac{1}{J}\sum_{j=1}^{J}\mathrm{ERB}\left(\Phi_{L,w}, \boldsymbol{X}_j\right)$.
   - $\overline{\mathrm{ERB}}(\Phi_{A,w}, \boldsymbol{X})$ and $\overline{\mathrm{ERB}}(\Phi_{H,w}, \boldsymbol{X})$ are analogously calculated.
   - $\mathrm{SD}(\mathrm{ERB}(\Phi_{L,w}, \boldsymbol{X})) = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}(\mathrm{ERB}(\Phi_{L,w}, \boldsymbol{X}_j) - \overline{\mathrm{ERB}}(\Phi_{L,w}, \boldsymbol{X}))^2}$.
   - $\mathrm{SD}(\mathrm{ERB}(\Phi_{A,w}, \boldsymbol{X}))$ and $\mathrm{SD}(\mathrm{ERB}(\Phi_{H,w}, \boldsymbol{X}))$ are analogously calculated.
   - $\overline{\mathrm{ERB}}\left(\widehat{\boldsymbol{\theta}_{L,w}}, \boldsymbol{X}\right) = \frac{1}{J}\sum_{j=1}^{J}\mathrm{ERB}\left(\widehat{\boldsymbol{\theta}_{L,w}}, \boldsymbol{X}_j\right)$.
   - $\overline{\mathrm{ERB}}(\widehat{\boldsymbol{\theta}_{A,w}}, \boldsymbol{X})$ and $\overline{\mathrm{ERB}}(\widehat{\boldsymbol{\theta}_{H,w}}, \boldsymbol{X})$ are analogously calculated.
   - $\mathrm{SD}\left(\mathrm{ERB}\left(\widehat{\boldsymbol{\theta}_{L,w}}, \boldsymbol{X}\right)\right) = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}(\mathrm{ERB}(\widehat{\boldsymbol{\theta}_{L,w}}, \boldsymbol{X}_j) - \overline{\mathrm{ERB}}(\widehat{\boldsymbol{\theta}_{L,w}}, \boldsymbol{X}))^2}$.
   - $\mathrm{SD}(\mathrm{ERB}(\widehat{\boldsymbol{\theta}_{A,w}}, \boldsymbol{X}))$ and $\mathrm{SD}(\mathrm{ERB}(\widehat{\boldsymbol{\theta}_{H,w}}, \boldsymbol{X}))$ are analogously calculated.
8. Obtain the average and standard deviation for the various performance measures of Step 5 across the J subsamples. Such numerical summaries are calculated in an intuitive manner analogous to the calculations of Step 7. Similarly, the average difference, or change, in transitioning from the pre- to post-fairness corrected threshold values (post-fairness corrected value minus pre-fairness corrected value), as well as the corresponding standard deviation, are also calculated in an intuitive manner analogous to the calculations in Step 7.

B.4.1. Visualizing variation across subsamples

To help understand the value of the additional resampling procedure detailed in Appendix B.4, for the identified "best" post-fairness corrected threshold values, this procedure was run using 200 random subsamples (i.e., $J = 200$). The resulting increase in Error Rate Balance at each threshold,

in transitioning from the pre- to post-fairness corrected threshold values, for each of these 200 subsamples is plotted in Fig. 4. The vertical blue line in each plot of this figure corresponds to no change as a result of the procedure, while anything to the left of this line corresponds to decreased fairness and anything to the right corresponds to increased fairness. From this figure, it is evident that for the average- and high-risk thresholds, all 200 subsamples result in a sizeable increase in fairness, whereas for the low-risk threshold, all but a few subsamples result in an improvement in fairness. However, from this plot, it is also evident that there is a non-marginal amount of variation in this improvement across subsamples, which must be accounted for when qualitatively and quantitatively assessing any improvement the procedure yields. Similar such plots and findings exists for alternative definitions of fairness and various predictive performance measures.

While the Error Rate Balance provides an aggregate measure of the parity in false positive and false negative rates across all pairings of protected attribute levels, it may also be of interest to assess those respective constituent measures in isolation. The procedure detailed in Appendix B.4 necessarily produces such values. For example, with the reunification algorithm, the average false negative and false positive rates, averaged across the same 200 random subsamples that yielded Fig. 4, for each level of the protected attribute at each threshold, for both the pre- and post-fairness corrected threshold values, are provided in Table 4.

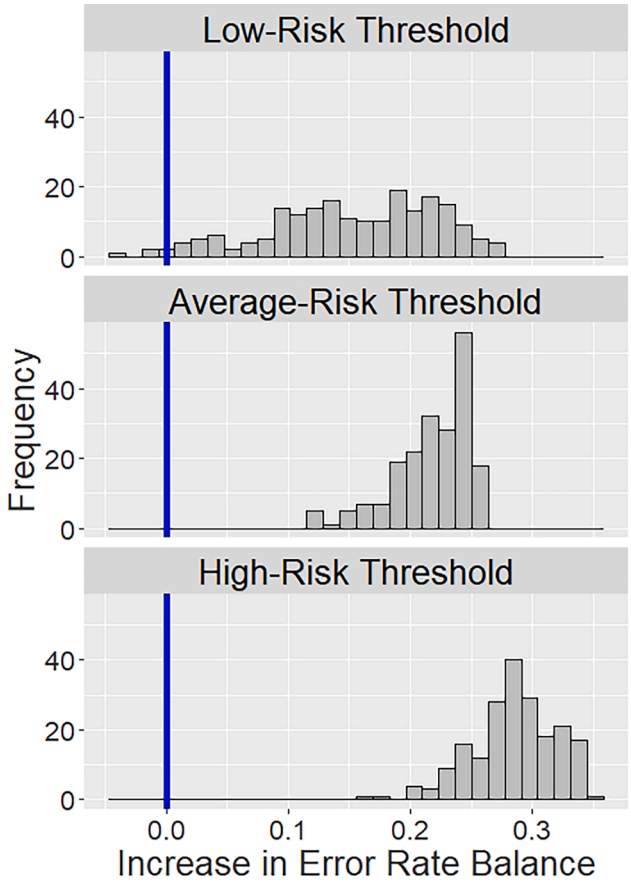


Fig. 4. The distribution of the increase in the extent to which Error Rate Balance is achieved, at each threshold for 200 random subsamples, in transitioning from the pre- to post-fairness corrected threshold values. The vertical blue line in each plot represents the point at which no change occurred as a result of the procedure, whereas anything to the left of this line corresponds to decreased fairness and anything to the right corresponds to increased fairness.

Table 4
The average false negative rate (FNR) and false positive rate (FPR), for both the pre-fairness-corrected (Pre-FC) and post-fairness-corrected (Post-FC) threshold values, across 200 random subsamples.

Threshold	Protected Attribute Level	FNR		FPR	
		Pre-FC	Post-FC	Pre-FC	Post-FC
Low-Risk	BL	0.113	0.129	0.722	0.697
Low-Risk	HPA	0.151	0.116	0.532	0.611
Low-Risk	NV	0.105	0.117	0.680	0.640
Low-Risk	WH	0.131	0.128	0.634	0.642
Average-Risk	BL	0.352	0.383	0.340	0.294
Average-Risk	HPA	0.381	0.324	0.206	0.252
Average-Risk	NV	0.325	0.343	0.310	0.287
Average-Risk	WH	0.390	0.381	0.279	0.287
High-Risk	BL	0.706	0.755	0.057	0.047
High-Risk	HPA	0.756	0.690	0.034	0.047
High-Risk	NV	0.669	0.683	0.055	0.046
High-Risk	WH	0.769	0.771	0.051	0.050

Appendix C. Demonstrating generalizability

Given that a single definition of algorithmic fairness is unlikely to be agreed upon across all Child Welfare jurisdictions, either within common use cases (e.g., hotline screening) or across distinct use-cases, there is a benefit to definition-agnostic "correction" procedures. To that end, the developed procedure can be applied using any of the nine group-level definitions of algorithmic fairness. To demonstrate this generalizability, the procedure was separately applied to three other definitions - Conditional Use Accuracy Equality, Treatment Equality, and Calibration.

C.1. Results under alternative definitions

The same specifications that were utilized with Error Rate Balance are used in applying the fairness correction procedure across each of Conditional Use Accuracy Equality, Treatment Equality, and Calibration. In particular, the procedure detailed in Appendix B.2 was performed under each of these three alternative definitions, utilizing 200 random subsamples (i.e., $I = 200$) and a grid of 101 w-values ranging between zero and one (i.e., $w = 0, 0.01, \ldots, 0.99, 1$). For each of these three definitions, for all 101 corresponding w-values, the procedure detailed in Appendix B.4 was then utilized to determine the extent to which the specified definition of fairness was achieved, on average, across 200 random subsamples (i.e., $J = 200$). For each of these three definitions, the "best" w-values across the three thresholds were then identified, where "best" corresponds to the post-fairness corrected threshold values yielding the greatest extent to which the specific definition of algorithmic fairness was achieved. Condensed results for each scenario are provided below to demonstrate the procedure's impact on the specified definition of algorithmic fairness.

C.1.1. Results under conditional use accuracy equality (CUAE)

The "best" w-values in this scenario are $w = 0.00, 0.01, 0.02$ for the low-risk, average-risk, and high-risk thresholds respectively. The average number of risk scores that changed in transitioning from the pre- to post-fairness corrected threshold values was 10.70% (standard deviation of 0.08%). For each threshold, the average change (post-fairness corrected minus pre-fairness corrected) in Conditional Use Accuracy Equality and in five common predictive performance measures as a result of transitioning from the pre- to post-fairness corrected threshold values is provided in Table 5, along with the corresponding standard deviations across the 200 random subsamples. From this table, two observations stand out. First, it is clear that the fairness-correction procedure has, at each threshold, meaningfully improved fairness with comparatively minimal cost to predictive performance. For example, at the low-risk threshold, the minimum parity in both the positive and negative predictive values between any two levels of the protected attribute (i.e., CUAE) has been increased, on average, by 0.111 (standard deviation of 0.019), with the cost of this improvement, on average, being no more than 0.045, in absolute value, to any of the overall predictive performance measures at that threshold. Importantly, the fairness at each threshold is not only improved, but notably good with an average Conditional Use Accuracy Equality of 0.95, 0.94, and 0.93 at the low-, average-, and high-risk thresholds, respectively. Second, even though the pre-fairness corrected risk scores are substantially fairer for CUAE than they are for Error Rate Balance, the correction procedure is still meaningfully able to increase the CUAE. In other words, the procedure was still able to substantively increase fairness even when the opportunity for improvement was comparatively less.

C.1.2. Results under treatment equality (TE)

The "best" w-values in this scenario are $w = 0.84, 0.74, 0.90$ for the low-risk, average-risk, and high-risk thresholds respectively. The average number of risk scores that changed in transitioning from the pre- to post-fairness corrected threshold values was 1.42% (standard deviation of 0.03%). For each threshold, the average change (post-fairness corrected minus pre-fairness corrected) in Treatment Equality and in five common predictive performance measures as a result of transitioning from the pre- to post-fairness corrected threshold values is provided in Table 6, along with the corresponding standard deviations across the 200 random subsamples. From this table, it appears that the fairness-correction procedure has, at each

Table 5
The average change (post-fairness corrected minus pre-fairness corrected), rounded to three decimal places, in Conditional Use Accuracy Equality (CUAE) and overall predictive performance measures at each threshold, along with corresponding standard deviations across 200 random subsamples. The predictive performance measures include Accuracy (ACC), False Negative Rate (FNR), False Positive Rate (FPR), Negative Predictive Value (NPV), and Positive Predictive Value (PPV).

Measure	Low-Risk Threshold Change		Average-Risk Threshold Change		High-Risk Threshold Change	
	Mean	SD	Mean	SD	Mean	SD
CUAE	0.111	0.019	0.045	0.021	0.111	0.026
ACC	0.033	0.001	0.013	0.001	0.000	< 0.0005
FNR	0.023	0.002	0.020	0.002	0.027	0.002
FPR	−0.045	0.001	−0.020	0.001	−0.006	< 0.0005
NPV	−0.003	0.001	−0.002	< 0.0005	−0.004	< 0.0005
PPV	0.009	< 0.0005	0.010	0.001	0.002	0.003

Table 6
The average change (post-fairness corrected minus pre-fairness corrected), rounded to three decimal places, in Treatment Equality (TE) and overall predictive performance measures at each threshold, along with corresponding standard deviations across 200 random subsamples. The predictive performance measures include Accuracy (ACC), False Negative Rate (FNR), False Positive Rate (FPR), Negative Predictive Value (NPV), and Positive Predictive Value (PPV).

Measure	Low-Risk Threshold Change		Average-Risk Threshold Change		High-Risk Threshold Change	
	Mean	SD	Mean	SD	Mean	SD
TE	0.072	0.066	0.079	0.056	0.095	0.049
ACC	−0.001	< 0.0005	−0.002	< 0.0005	0.001	< 0.0005
FNR	0.000	0.001	−0.003	0.001	−0.001	0.001
FPR	0.001	< 0.0005	0.003	< 0.0005	−0.000	< 0.0005
NPV	−0.000	< 0.0005	0.000	< 0.0005	0.000	< 0.0005
PPV	−0.000	< 0.0005	−0.001	< 0.0005	0.003	0.001

Table 7
The average change (post-fairness corrected minus pre-fairness corrected), rounded to three decimal places, in Calibration (CAL) at each Risk Score, along with corresponding standard deviations across 200 random subsamples.

Measure	**Change in Risk Score** $S1$		**Change in Risk Score** $S2$		**Change in Risk Score** $S3$		**Change in Risk Score** $S4$	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CAL	0.185	0.051	0.154	0.064	0.020	0.0271	0.021	0.006

Table 8
The average change (post-fairness corrected minus pre-fairness corrected), rounded to three decimal places, in overall predictive performance measures at each threshold, along with corresponding standard deviations across 200 random subsamples, after fairness correcting under Calibration (CAL). The predictive performance measures include Accuracy (ACC), False Negative Rate (FNR), False Positive Rate (FPR), Negative Predictive Value (NPV), and Positive Predictive Value (PPV).

Measure	**Low-Risk Threshold Change**		**Average-Risk Threshold Change**		**High-Risk Threshold Change**	
	Mean	SD	Mean	SD	Mean	SD
ACC	0.032	0.001	−0.000	< 0.0005	0.000	< 0.0005
FNR	0.016	0.002	−0.000	0.001	0.006	0.001
FPR	−0.042	< 0.0005	0.000	< 0.0005	−0.001	< 0.0005
NPV	−0.001	0.001	−0.000	< 0.0005	−0.001	< 0.0005
PPV	0.009	< 0.0005	−0.000	< 0.0005	0.001	0.001

threshold, marginally to meaningfully improved fairness with essentially no cost to predictive performance. For example, at the high-risk threshold, the minimum parity in the ratio of false positives to false negatives between any two levels of the protected attribute (i.e., TE) has been increased, on average, by 0.095 (standard deviation of 0.049), with the cost of this improvement, on average, being no more than 0.003, in absolute value, to any of the overall predictive performance measures at that threshold. Importantly, the fairness at each threshold is not only improved, but categorically good with an average Conditional Use Accuracy Equality of 0.86, 0.91, and 0.91 at the low-, average-, and high-risk thresholds, respectively.

C.1.3. Results under calibration (CAL)

The "best" w-values in this scenario were the same, $w = 0.05$, at all three thresholds since a single penalized optimization problem must be solved across all three thresholds simultaneously. The average number of risk scores that changed in transitioning from the pre- to post-fairness corrected threshold values was 5.57% (standard deviation of 0.05%). For each risk score, the average change (post-fairness corrected minus pre-fairness corrected) in Calibration as a result of transitioning from the pre- to post-fairness corrected threshold values is provided in Table 7, along with the corresponding standard deviations across the 200 random subsamples. Similarly, for each threshold, the average change (post-fairness corrected minus pre-fairness corrected) in five common predictive performance measures as a result of transitioning from the pre- to post-fairness corrected threshold values is provided in Table 8, along with the corresponding standard deviations across the 200 random subsamples. From these tables, two observations again stand out. First, the fairness-correction procedure has, at three of the fours risk scores, substantively improved fairness with comparatively minimal cost to predictive performance across the three thresholds. For example, at the $S1$ risk score, the minimum parity in the proportion of child-reunification pairs that ultimately return to substitute care between any two levels of the protected attribute (i.e., CAL) has been increased, on average, by 0.185 (standard deviation of 0.051); the cost of this improvement, on average, is no more than 0.042, in absolute value, to any of the overall predictive performance measures across all three thresholds. Importantly, the fairness of the risk scores is not only improved, but relatively good with an average Calibration of 0.71, 0.89, 0.93, and 0.84 across the four scores, respectively. Second, despite Calibration being subtly, but importantly, different in how it assesses fairness, the utilized correction procedure is still effective at increasing fairness according to this definition.

C.2. Some final thoughts on generalizability

While the demonstrated generalizability of the developed correction procedure does not preclude the complex and difficult conversations surrounding the choice of definition of algorithmic fairness, it does reveal a robust approach for increasing fairness once that decision has been made. Furthermore, although not demonstrated here due to small sample sizes, for larger jurisdictions the procedure is theoretically applicable across an arbitrary number of protected attributes and thresholds. Such functionality, among other things, enables the utilization of a multidimensional (e.g., race, gender, and disability status) protected attribute with a finely-grained risk scoring system.

Appendix D. Child-transition features for machine learning classifier

Table 9 lists the features available for each child-transition observation, constructed from administrative data. Features described with bracketed terms (e.g., # Days) represent multiple which vary over a set of options, such as the number of days into the past, or the type of allegation named in a report of abuse/neglect. These sets of options will vary between jurisdiction. Careful consideration must be made in selecting and defining appropriate features in order to ensure the features represent their intended constructs in a valid and reliable fashion.

Table 9
List of child-transition features used in the machine learning classifier.

Feature Name	Description
AgeAtStart	Current age of child
Gender_Male	Child is male
nuPERPS_roleVCT	# prior alleged perps associated with the child
nuRPT_roleVCT_FM	# prior reports for which child was the victim
nuPERPS_roleVCT_FM	# prior alleged perps associated with the child
SP_n_{# Days}	# service placements in past {30,60,90,180,365 Days}
SP_n_{Placement Type}_{# Days}	# service placements in past {# Days} of {Placement Type Category}
SP_n_InHome_{# Days}	# In–Home service episodes in past {# Days}
INV_FM_n_{# Days}	# investigations in past {# Days}
INV_FM_n_Safety_Decision_{# Days}	# investigations in past {# Days} with Safety Decision of Unsafe
INV_FM_n_Founded_{Allegation Type}_{# Days}	# founded investigations in past {# Days} with {Allegation}
INV_FM_n_UTD_{Allegation Type}_{# Days}	# undetermined investigations in past {# Days} with {Allegation}
INV_FM_n_Unfounded_{Allegation Type}_{# Days}	# unfounded investigations in past {# Days} with {Allegation}
INV_FM_n_NotSDU_{Allegation Type}_{# Days}	# investigations in past {# Days} with {Allegation}
INV_NONFM_n_{# Days}	# non-familial investigations in past {# Days}
RPT_FM_n_Assigned_{# Days}	# assigned reports involving the child in past {# Days}
RPT_FM_n_{# Days}	# reports involving the child in past {# Days}
RPT_FM_n_Role_Victim_{# Days}	# reports involving the child as victrim in past {# Days}
RPT_FM_n_Role_Perp_{# Days}	# reports involving the child as perp in past {# Days}
RPT_FM_n_Assigned_{Allegation Type}_{# Days}	# assigned reports involving the child in past {# Days} with {Allegation}
RPT_FM_n_CAS_{Allegation Type}_{# Days}	# closed reports involving child in past {# Days} with {Allegation}
RPT_NONFM_n_{# Days}	# non-familial referrals in past {# Days}
{Parent}_RPT_asPERP_n_Assigned_{# Days}	# assigned reports involving {Parent} in past {# Days}
{Parent}_RPT_asPERP_n_{# Days}	# reports involving {Parent} in past {# Days}
{Parent}_RPT_asPERP_n_Assigned_{Allegation Type}_{# Days}	# assigned reports involving {Parent} in past {# Days} with {Allegation}
{Parent}_RPT_asPERP_n_CAS_{Allegation Type}_{# Days}	# reports involving {Parent} in past {# Days}
{Parent}_INV_asPERP_n_{# Days}	# investigations involving {Parent} in past {# Days}
{Parent}_INV_asPERP_n_Founded_{Allegation Type}_{# Days}	# investigations involving {Parent} in past {# Days} with {Allegation}
{Parent}_INV_asPERP_n_Safety_Dec_Unsafe_{# Days}	# unsafe determinations involving {Parent} in past {# Days}
{Parent}_INV_asPERP_n_Children_RemHome_{# Days}	# investigations involving {Parent} in past {# Days} without removal
RECENT_RPT_Assigned	Child's most recent report was assigned
RECENT_RPT_Assigned_{Allegation Type}	Child's most recent report was assigned and involved {Allegation}
RECENT_RPT_CAS_{Allegation Type}	Child's most recent report was closed and involved {Allegation}
RECENT_RPT_Familial	Child's most recent report was familial
RECENT_RPT_MandatoryReporter	Child's most recent report was made by a mandatory reporter
RECENT_RPT_Source_{Reporter Category}	Child's most recent report was made by {Reporter Category}
RECENT_RPT_Role_n_{Role}	# named on child's recent report with a given {Role} designation
PA_CAT	Child's protected attribute level

References

Ainsworth, F., & Maluccio, A. N. (1998). The policy and practice of family reunification. *Australian Social Work, 51*(1), 3–7.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L., (2016). *Machine bias: There's software used across the country to predict future criminals and it's biased against blacks*. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2020-09-28.

Oregon Department of Human Services. (2015). *Oregon Child and Family Services Plan*. URL https://www.oregon.gov/dhs/children/Pages/data-publications.aspx Accessed: 2020-09-28.

Oregon Department of Human Services. (2014). *Service Equity Framework*. URL https://www.oregon.gov/DHS/SENIORS-DISABILITIES/SUA/AAABusinessTraining/Service Equity presented April 2014.pdf. Accessed: 2020-09-28.

Office of Reporting Research Analytics and Implementation, Oregon Department of Human Services. (2019). *Safety at Screening Tool Development and Execution Report*. URL https://www.oregon.gov/DHS/ORRAI/Pages/index.aspx. Accessed: 2020-09-28.

Purdy, J., Glass, B., & Pakseresht, F. (2018). *Fairness in Machine-Learning-Generated Risk Scores via Equitable Thresholding*. URL https://www.oregon.gov/DHS/ORRAI/Pages/index.aspx. Accessed: 2020-09-28.

Barocas, S., Hardt, & M., Narayanan, A., (2018). *Fairness and machine learning*. URL https://www.fairmlbook.org. Accessed: 2021-03-17.

Bechavod, Y., & Ligett, K., (2017). *Learning fair classifiers: A regularization-inspired approach*. arXiv preprint arXiv:1707.00044v2.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A., (2017). Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*.

Biehal, N. (2007). Reuniting children with their families: Reconsidering the evidence on timing, contact and outcomes. *British Journal of Social Work, 37*(5), 807–823.

Biehal, N., Sinclair, I., & Wade, J. (2015). Reunifying abused or neglected children: Decision-making and outcomes. *Child Abuse & Neglect, 49*, 107–118.

Binns, R., (2019). On the apparent conflict between individual and group fairness. *arXiv preprint arXiv:1912.06883v1*.

Bower, A., Kitchen, S.N., Niss, L., Strauss, M.J., Vargo, A., & Venkatasubramanian, S., (2017). Fair pipelines. *arXiv preprint arXiv:1707.00391v1*.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Buolamwini, J., & Gebru, T., (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceeding of the 2018 conference on fairness, accountability and transparency* (pp. 1–15).

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., & Smith, A., (2019). From soft classifiers to hard decisions: How fair can we be? In *Conference on fairness, accountability, and transparency*.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y., (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1-4.

Chouldechova, A., (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R., (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 2018 conference on fairness, accountability and transparency* (pp. 134–148).

Chouldechova, A., Putnam-Hornstein, E., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018b). A case study of algorithm-assisted decision making in cild maltreatment hotline screening decisions. *Proceedings of Machine Learning Research, 81*, 1–15.

Chouldechova, A., & Roth, A., (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810v1*.

Coston, A., Mishler, A., Kennedy, E.H., & Chouldechova, A., (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 582–593).

Dheeru, D., & Karra Taniskidou, E., (2017). *UCI machine learning repository*. URL http://archive.ics.uci.edu/ml. Accessed: 2020-09-28.

DHHS, U.S., (2016). *Comprehensive child welfare information system; final rule*. URL https://www.govinfo.gov/content/pkg/FR-2016-06-02/pdf/2016-12509.pdf. Accessed: 2020-09-28.

Drake, B., Jonson-Reid, M., Ocampo, M. G., Morrison, M., & Dvalishvili, D. (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.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*(1), 1–24.

Ensign, D., Friedler, S., Neville, S., Scheidegger, C., & Venkatasubramanian, S., (2018). Runaway feedback loops in predictive policing. In *Proceedings machine learning research. conference on fairness, accountability, and transparency* (pp. 1–12).

Esposito, T., Delaye, A., Chabot, M., Trocmé, N., Rothwell, D., Hélie, S., & Robichaud, M. J. (2017). The effects of socioeconomic vulnerability, psychosocial services, and social service spending on family reunification: A multilevel longitudinal analysis. *International Journal of Environmental Research and Public Health, 14*(9), 1040.

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., & Roth, D., (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 329–338).

Green, B., (2020). The false promise of risk assessments: Epistemic reform and the limits of fairness. In *Conference on fairness, accountability and transparency*.

Green, B., & Chen, Y., (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 90–99).

Hardt, M., Price, E., & Srebro, N., (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413v1*.

Keddell, E. (2019). Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice. *Social Sciences, 8*(10), 281.

Kleinberg, J., Mullainathan, S., & Raghavan, M., (2016). Inherent trade-offs in the fair determinatinon of risk scores. *arXiv preprint arXiv:1609.05807v2*.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S., (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 275–284).

Lipton, Z.C., Chouldechova, A., & McAuley, J., (2019). Does mitigating ml's disparate impact require disparate treatment? *arXiv preprint arXiv:1711.07076v3*.

Mitchell, S., Potash, E., Barocas, S., & Alexander D'Amour, K.L., (2020). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867v3*.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*, 447–453.

Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170359.

Passi, S., & Barocas, S., (2019). Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 39–48).

Pedreshi, D., Ruggieri, S., & Turini, F., (2008). Discrimination-aware data mining. In *14th acm sigkdd*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *31st conference on neural information processing systems (nips 2017), Long Beach, CA, USA*.

Samant, A., Horowitz, A., Xu, K., & Beiers, S., (2021). *Family surveillance by algorithm: The rapidly spreading tools few have heard of*. https://www.aclu.org/sites/default/files/field_document/2021.09.28a_family_surveillance_by_algorithm.pdf. Accessed: 2022-05-26.

Terling, T. (1999). The efficacy of family reunification practices: Reentry rates and correlates of reentry for abused and neglected children reunited with their families. *Child Abuse & Neglect, 23*(12), 1359–1370.

Veale, M., Van Kleek, M., & Binns, R., (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–14).

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Fairware' 18: Ieee/acm international workshop on software fairness, gothenburg, sweden*.

Zafar, M.B., Valera, I., Rodriguez, M.G., & Gummadi, K.P., (2017). Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.