Enhancing a Recidivism Prediction Tool With Machine Learning: Effectiveness and Algorithmic Fairness

Marzieh Karimi-Haghighi
Universitat Pompeu Fabra
marzieh.karimihaghighi@upf.edu

Carlos Castillo
Universitat Pompeu Fabra
carlos.castillo@upf.edu

ABSTRACT

This paper addresses a key application of Machine Learning (ML) in the legal domain, studying how ML may be used to increase the effectiveness of a criminal recidivism risk assessment tool named RisCanvi, without introducing undue biases. The two key dimensions of this analysis are predictive accuracy and algorithmic fairness. ML-based prediction models obtained in this study are more accurate at predicting criminal recidivism than the manually-created formula used in RisCanvi, achieving an AUC of 0.76 and 0.73 in predicting violent and general recidivism respectively. However, the improvements are small, and it is noticed that algorithmic discrimination can easily be introduced between groups such as national vs foreigner, or young vs old. It is described how effectiveness and algorithmic fairness objectives can be balanced, applying a method in which a single error disparity in terms of generalized false positive rate is minimized, while calibration is maintained across groups. Obtained results show that this bias mitigation procedure can substantially reduce generalized false positive rate disparities across multiple groups. Based on these results, it is proposed that ML-based criminal recidivism risk prediction should not be introduced without applying algorithmic bias mitigation procedures.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**.

KEYWORDS

criminal recidivism, risk assessment, algorithmic fairness

ACM Reference Format:
Marzieh Karimi-Haghighi and Carlos Castillo. 2021. Enhancing a Recidivism Prediction Tool With Machine Learning: Effectiveness and Algorithmic Fairness. In *Eighteenth International Conference for Artificial Intelligence and Law (ICAIL'21), June 21–25, 2021, São Paulo, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3462757.3466150

1 INTRODUCTION

Risk assessment is a necessary process in many important decisions such as public health, information security, project management, auditing, and criminal justice. Since the 1920s, violence risk assessment tools have been progressively used in criminal justice by probation and parole officers, police, and psychologists to assess the risk of harm, sexual, criminal, and violent offending in more than 44 countries [22, 32]. In comparison to traditional prediction methods and unstructured clinical judgments, risk assessment tools offer superior accuracy and performance [18]. In this regard, factors such as the availability of large databases, inexpensive computing power, and developments in statistics and computer science have brought an increase in the accuracy and applicability of these structured tools [3]. Such advances have effectively increased the use of tools based on Machine Learning (ML) in criminal justice decisions for risk forecasting [4, 7, 8]. Today, various semi-structured protocols for assessing risk of recidivism can be found in different countries including the U.S. [16], the U.K. [21], Canada [24], Austria [30], and Germany [13]. In Spain, among current violence risk assessment tools including SAVRY, PCL-R, HCR-20, SVR-20, and SARA, RisCanvi is a relatively new tool for risk assessment of recidivism. It was originally developed in 2009 in response to concerns of Catalan prison system officials regarding violent recidivism among offenders after their sentences.

Research contribution. In this study, the effectiveness and algorithmic fairness of RisCanvi risk assessment tool are evaluated in comparison to ML models such as logistic regression, perceptron, and support-vector machines, in violent and general recidivism prediction. The effectiveness of the ML models are evaluated and compared to RisCanvi in terms of various metrics including AUC, Generalized False Positive (GFPR), and Generalized False Negative (GFNR). Also, potential algorithmic bias introduced by the ML methods is evaluated in both violent and general recidivism prediction. Given that model learning may lead to unfairness [11, 12, 34], the impact of the obtained ML models is compared along nationality (national origin vs foreign origin) and age (young vs old). Then some differences are addressed through a mitigation procedure [29], which try to equalize GFPR across nationality and age groups while preserving the calibration in each group.

The rest of this paper is organized as follows. Section 2 outlines related work. In Section 3, the RisCanvi risk assessment tool and the dataset used in this study are described. The methodology including the ML models and algorithmic fairness analysis are presented in Section 4. Results are given in Section 5, and a procedure to mitigate algorithmic discrimination is used in Section 6. Finally, the results are discussed and the paper is concluded in Section 7.

2 RELATED WORK

The introduction of algorithms for risk assessment in criminal justice is a controversial topic, and perhaps one that has motivated a great deal of research on algorithmic fairness.

ICAIL'21, June 21–25, 2021, São Paulo, Brazil

ACM ISBN 978-1-4503-8526-8/21/06…$15.00
https://doi.org/10.1145/3462757.3466150

In seminal research done by investigative journalism organization ProPublica [2, 25] it was concluded that a widely-used program named Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) is biased against African American defendants. A follow-up study [19] found that COMPAS outcomes systematically over-predict risk for women, thereby indicating systemic gender bias. However, the findings of the ProPublica study were rejected by Northpointe (COMPAS developer), claiming their algorithm is fair because it is well calibrated [17]. Moreover, in this report it is shown that the COMPAS risk scales exhibit accuracy equity and predictive parity.

In contrast to the case of COMPAS, other studies have shown that other risk assessment tools such as the Post Conviction Risk Assessment (PCRA), the Structured Assessment of Violence Risk in Youth (SAVRY) and the Youth Level of Service/Case Management Inventory (YLS/CMI) do not exhibit racial bias in the recidivism prediction [28, 33]. In a more recent study focused on SAVRY [26, 34], it is shown that although machine learning models could be more accurate than the simple summation used to compute SAVRY scores, they would introduce discrimination against some groups of defendants.

There are many different definitions of algorithmic fairness [27], some of which are incompatible with one another. It is impossible to satisfy all of them simultaneously except in pathological cases (such as a perfect classifier), and in general it is impossible to maximize algorithmic fairness and accuracy at the same time [5, 6]. Hence, there are necessary trade-offs between different metrics [6, 10, 23]. In this regard, some studies [20, 36, 37] try to mitigate potential algorithmic discrimination by satisfying equalized odds or in other words avoiding disparate mistreatment along different sensitive groups. In addition, due to the importance of the calibration in risk assessment tools [6, 17], some previous work has also tried to minimize error disparity across groups while maintaining calibrated probability estimates [29].

The most closely related previous work is Pleiss et al. [29], where algorithmic bias in a machine learned risk assessment (COMPAS) is minimized by equalizing generalized false positive rates along different races, finding this equalization to be incompatible with calibration. In contrast, in the work presented on this paper, we start from an expert-based risk assessment method, which is not machine learned, and propose a new machine learning model to replace it, describing the effects of algorithmic bias mitigation on both the original and the machine learned model. Additionally, we find that in RisCanvi equalization along nationality and age groups is not entirely incompatible with calibration.

3 RISCANVI DATASET

3.1 The RisCanvi Risk Assessment Tool

RisCanvi was introduced as a multi-level risk assessment protocol for violence prevention in the Catalan prison system in 2009 [1]. This protocol is applied multiple times during an inmate's period in prison; the official recommendation is to do so every six months or at the discretion of the case manager. RisCanvi is not a questionnaire. Instead, each inmate is interviewed by professionals. In the original RisCanvi protocol, risk is determined for each inmate relative to four possible outcomes: self-directed violence, violence in the prison facilities, committing further violent offenses, and breaking prison permits. A fifth risk score was introduced more recently for general recidivism [31].

Two versions of the RisCanvi protocol were created, an abbreviated one of 10 items for screening (RisCanvi-S), and a complete one of 43 items (RisCanvi-C). Risk items can be categorized into five different categories: Criminal/Penitentiary, Biographical, Family/Social, Clinical, Attitudes/ Personality. These items can also be divided into static factors (such as "criminal history of family" and "age of starting violent activity") and dynamic factors (such as "member of socially vulnerable groups" and "pro-criminal or antisocial attitudes").

3.2 Dataset

The anonymized dataset used on this research comprises 7,239 offenders who first entered the prison between 1989 and 2012 and who were evaluated with the RisCanvi protocol between 2010 and 2013. Only offenders for which nationality information was recorded were kept that comprises 2,634 offenders. The result population was filtered in terms of their violent/general recidivism, freedom and last RisCanvi evaluation dates considering the following conditions: inmates who were released at most 9 months after their last RisCanvi evaluation, and for which violent/general recidivism (or its absence) was recorded at most two years after their release. Finally, samples with the size of 2,027 (out of 2,634) were reached. Among this population, 146 committed a violent offence (violent recidivism) and 310 committed a violent or non-violent offence (general recidivism) after being released. The data includes all of the information for the two RisCanvi versions (RisCanvi-S and RisCanvi-C). This study is focused on the RisCanvi-C protocol which is the complete version done after RisCanvi-S and it consists of more risk factors which results in three risk levels (low, medium, and high).

3.3 Violent and General Recidivism

This work is focused on RisCanvi protocol to assess Violent Recidivism ("REVI" in the RisCanvi manual) and General Recidivism ("REGE" in the RisCanvi manual) risks in sentenced inmates. REVI and REGE risks are outcomes predicted using two different sub-sets of risk factors. REVI risk is obtained using 23 items out of the 43 risk factors of the RisCanvi-C version plus two demographic features (gender and nationality) and to compute REGE risk, 14 items (out of 43 risk factors of the RisCanvi-C version) are used. In RisCanvi-C, each of the REVI and REGE scores has been computed by applying the summation of their related features in a hand-crafted formula, then using two cut-offs, obtaining three risk levels (details in [1]).

The distribution of REVI and REGE risk scores in the last RisCanvi evaluation is compared by nationality and age groups. Grouping by gender is not considered as the number of women in the sample is too small to draw robust conclusions. The comparison shows that recidivism risk scores have approximately similar distributions along nationality and age group except for the REVI score in nationality group which shows that foreigners tend to have lower REVI risk scores compared to Spaniards (Figures are omitted for brevity). For age groups, 30 years old is used as a cut-off, as criminology research suggests that the types of offense and context are different

for people under 30 and over 30 (see, e.g., [35]). This age is also used as a cut-off for young and old people in the design of the RisCanvi protocol. In the present dataset, the majority of the population are Spanish nationals (70%) and older than 30 years old (74%).

According to the average violent and general recidivism rates for nationality and age groups, it can be seen that in general, foreigners and older offenders have a lower recidivism rate.

4 METHODOLOGY

The goal of this study is to compare the effectiveness and fairness of Machine Learning (ML) models and the RisCanvi risk assessment tool in the prediction of violent and general recidivism.

4.1 ML-based Models

Different ML methods, such as logistic regression, multi-layer perceptron (MLP), and support vector machines (SVM) are used. The ground truth is the violent/general recidivism, which is recorded at most two years after the inmate's release.

Different sub-sets of features are tested as input to the ML models, such as 43 RisCanvi-C items, Violent Recidivism (REVI)/General Recidivism (REGE) risk items, and a set of features selected from 43 risk items using a feature selection method. In addition, three demographic features (gender, nationality, and age) are used as general input features. Finally, the average of REVI/REGE risk scores over all of the RisCanvi evaluations from the first to the last evaluation is added.

The split of the two sets is done k times using stratified k-fold cross-validation, reporting average results.

4.2 Algorithmic Fairness

Algorithmic fairness is evaluated by comparing the impact of the risk prediction method across nationality and age groups.

As it is known, model calibration is a necessary condition, especially in criminal justice risk assessments [6, 17]. If the risk tool is not calibrated with respect to different groups, then the same risk estimate carries different meanings and cannot be interpreted equally for different groups. Furthermore, creating parity in the error rates of different groups ("equalized odds") is a well-established method to mitigate algorithmic discrimination in automatic classification. Previous work has also emphasized the importance of this algorithmic fairness metric for this particular application [20, 36, 37]. Hence, to mitigate potential algorithmic discrimination, a relaxation method [29] is used in this paper which seeks to satisfy equalized odds or parity in the error rates (generalized false positive rate and generalized false negative rate) while preserving calibration in each sub-group of nationality and age. In most cases, calibration and equalized odds are mutually incompatible goals [10, 23], so in this method it is sought to minimize only a single error disparity across groups while maintaining calibration probability estimates.

Generalized False Positive Rate (GFPR) and Generalized False Negative Rate (GFNR) are the standard notions of false-positive and false-negative rates that are generalized for use with probabilistic classifiers [29]. If variable x represent an inmate's features vector, y indicates whether or not the inmate recidivists, G_1, G_2 are the two different groups, and h_1, h_2 are binary classifiers which classify samples from G_1, G_2 respectively, GFPR and GFNR are defined as follows [29]: the GFPR of classifier h_t for group G_t is $c_{fp}(h_t) = \mathbb{E}_{(x,y)\sim G_t}[h_t(x) \mid y = 0]$. GFPR is the average probability of being recidivist that the classifier estimates for people who actually do not recidivate. Conversely, the GFNR of classifier h_t is $c_{fn}(h_t) = \mathbb{E}_{(x,y)\sim G_t}[(1-h_t(x)) \mid y = 1]$. So the two classifiers h_1 and h_2 show probabilistic equalized odds across groups G_1 and G_2 if $c_{fp}(h_1) = c_{fp}(h_2)$ and $c_{fn}(h_1) = c_{fn}(h_2)$.

Classifier h_t is said to be *well-calibrated* if $\forall p \in [0, 1]$, $\mathrm{P}_{(x,y)\sim G_t}[y = 1 \mid h_t(x) = p] = p$. To prevent the probability scores from carrying group-specific information, both classifiers h_1 and h_2 are calibrated with respect to groups G_1 and G_2 [6, 17].

Table 1: Effectiveness of models in violent and general recidivism prediction

Risk	Violent Recidivism			General Recidivism		
Model	AUC	GFNR	GFPR	AUC	GFNR	GFPR
LR	0.76	0.82	0.06	0.73	0.75	0.14
RisCanvi_score	0.72	0.87	0.07	0.70	0.79	0.14

5 RESULTS

5.1 Effectiveness Evaluation

Among logistic regression (LR), multi-layer perceptron (MLP) and support vector machines, the best results were obtained using LR for both violent and general recidivism predictions. Hence, the non-LR based models are omitted for brevity. The final set of features used for the model consists of a sub-set of the 43 risk items of the RisCanvi evaluation selected using a feature selection method (based on a linear model with L1-based penalization to yield sparse coefficients), the average Violent Recidivism (REVI)/General Recidivism (REGE) score (from the first to the last RisCanvi evaluation), gender, nationality, and age at the time of the last evaluation.

Results in terms of AUC-ROC, GFNR, and GFPR are presented and compared with the existing RisCanvi protocol in Table 1 for both violent and general recidivism prediction. These results are compared against RisCanvi_score, which is a number resulting from the application of the RisCanvi formula.

In both violent and general recidivism prediction, LR yields better results than RisCanvi in terms of all metrics. However, the results are close to RisCanvi. In general, the LR model is more accurate than RisCanvi, although by a small amount, which is surprising considering that RisCanvi was not computationally optimized for predictive accuracy.

5.2 Algorithmic Fairness Evaluation

The results for the analysis of algorithmic fairness in all metrics along nationality (national and foreigner), and age groups (young and old inmates) are shown in Table 2 for violent and general recidivism prediction. In the LR_calibrated model, the predictions have been calibrated with respect to each of the two sub-groups in nationality and age.

For violent recidivism, all models show a bias against nationals in terms of GFPR. The difference is less noticeable in RisCanvi. In

Table 2: Effectiveness of models in violent and general recidivism prediction per group

Risk	Violent Recidivism									General Recidivism								
Model	LR			LR_Calibrated			RisCanvi			LR			LR_Calibrated			RisCanvi		
Group/Metrics	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR
National	0.81	0.77	0.07	0.81	0.81	0.06	0.76	0.85	0.08	0.78	0.70	0.15	0.77	0.73	0.13	0.72	0.78	0.14
Foreigner	0.85	0.87	0.05	0.85	0.85	0.04	0.72	0.91	0.05	0.68	0.80	0.11	0.72	0.77	0.11	0.59	0.83	0.13
$\frac{\text{National}}{\text{Foreigner}}$ (Ratio)	(0.95)	(0.88)	(1.64)	(0.95)	(0.95)	(1.50)	(1.05)	(0.93)	(1.44)	(1.14)	(0.87)	(1.30)	(1.07)	(0.95)	(1.20)	(1.22)	(0.94)	(1.08)
Young	0.84	0.78	0.08	0.84	0.83	0.06	0.79	0.86	0.07	0.67	0.74	0.17	0.72	0.75	0.15	0.58	0.82	0.14
Old	0.83	0.78	0.06	0.83	0.83	0.06	0.76	0.85	0.07	0.78	0.71	0.12	0.75	0.74	0.11	0.75	0.78	0.14
$\frac{\text{Young}}{\text{Old}}$ (Ratio)	(1.02)	(1.00)	(1.26)	(1.02)	(1.01)	(1.11)	(1.04)	(1.00)	(1.03)	(0.85)	(1.04)	(1.38)	(0.96)	(1.01)	(1.37)	(0.77)	(1.06)	(1.03)

LR model, we can also observe higher GFPR for young inmates compared to old offenders. In general, LR_calibrated and RisCanvi models lead to more algorithmically fair results along both nationality and age in terms of all metrics, except for the metrics in which all the models show discrimination.

The results for general recidivism prediction show higher AUC for nationals compared to foreigners in RisCanvi. In terms of GFPR, the LR and LR_calibrated models show discrimination against national group. In age group, LR and LR_calibrated models show higher GFPR along young compared to old group. In terms of AUC, we can see more discrimination against young inmates in RisCanvi compared to other models. As a result, LR_calibrated model shows better algorithmic fairness properties across nationality and more balanced values can be observed along age group in RisCanvi.

6 EQUALIZED ODDS AND CALIBRATION

In this section, it is tried to achieve parity along nationality and age groups in terms of two fairness metrics simultaneously. For this purpose, the method introduced by Pleiss et al. [29] is used that seeks parity in Generalized False Positive Rate (GFPR) or Generalized False Negative Rate (GFNR) while preserving calibration in each sub-group of nationality and age. The conclusion from the previous section based on the results obtained per group in Table 2, is that in both violent and general recidivism predictions, machine learning models show inequality in terms of GFPR along nationality and age. RisCanvi also shows an imbalance in GFPR values along nationality groups in violent recidivism prediction.

Hence, it is tried to create parity in this metric while preserving calibration in each group. The results after bias mitigation is presented in Table 3 for violent and general recidivism prediction. The obtained models are referred to in the following as LR-Equalized, LR_Calibrated-Equalized, and RisCanvi-Equalized.

By comparing the results before and after this bias mitigation (Table 2 and Table 3 respectively) in violent recidivism, it can be seen that the discrimination in GFPR has decreased in the order of 0.08-0.26 and 0.06-0.09 along nationality and age groups respectively. Also, comparing the results before and after bias mitigation in general recidivism shows that there are reductions in GFPR disparity in the orders of 0.03-0.04 and 0.16-0.19 along nationality and age groups respectively. However, in both violent and general recidivism prediction, the decline in GFPR bias is obtained at the expense of further inequalities in other metrics.

7 DISCUSSION AND CONCLUSIONS

The effectiveness and fairness of Machine Learning (ML) models in violent and general recidivism prediction were compared to the RisCanvi risk assessment tool, an in-use model created by experts. ML models were generated with AUC of 0.76 and 0.73 in violent and general recidivism prediction respectively which shows slightly better results compared to the AUC of RisCanvi protocol which is 0.72 and 0.70. It is noteworthy that in this type of task, predictions are not very accurate in general (existing recidivism prediction tools typically have AUC in the range of 0.57-0.74 [9, 14, 15]), and it is found that a hand-crafted formula created by experts is quite comparable to a machine-learned one. Although the improvement in accuracy by ML would be insufficient on its own to support its introduction as a risk assessment tool, a key element of ML models is their flexibility. An ML model can be re-trained with newer data, and incorporate new factors as the population of inmates changes and more data on recidivism becomes available.

By studying differential treatment of RisCanvi and ML models across different groups, it can be stated that depending on the desired metric and groups, machine learning and human expert can lead to different but comparable results. An advantage of ML models is that the emphasis on different metrics can be changed during the modeling as legal or policy changes are introduced. In this study, results in Table 2 showed that in both violent and general recidivism predictions, there is an inequality in terms of Generalized False Positive Rate (GFPR) metric along nationality and age groups. So using a relaxation method [29], it was tried to set parity in GFPR while preserving calibration in each sub-group of nationality and age. The results after bias mitigation (in Table 3) showed that GFPR disparity in violent and general recidivism has been respectively decreased at most 0.26 and 0.04 along nationality and 0.09 and 0.19 along age, however, in exchange for inequalities in some other metrics.

A robust conclusion from this work is that in a context in which predictive factors neither determine nor yield a clear signal of low/medium/high recidivism risk, ML cannot be considered a silver bullet. At the very least, improvements in accuracy need to be carefully contrasted with potential issues of algorithmic fairness

Table 3: Equalized GFPR while preserving calibration in violent and general recidivism prediction

Risk	Violent Recidivism									General Recidivism					
Model	LR-Equalized			LR_Calib-Equalized			RisCanvi-Equalized			LR-Equalized			LR_Calib-Equalized		
Group/Metrics	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR	AUC	GFNR	GFPR
National	0.81	0.77	0.07	0.81	0.81	0.06	0.76	0.85	0.08	0.78	0.70	0.15	0.67	0.78	0.14
Foreigner	0.64	0.92	0.05	0.61	0.91	0.05	0.62	0.92	0.06	0.61	0.81	0.12	0.53	0.88	0.12
$\frac{\text{National}}{\text{Foreigner}}$ (Ratio)	(1.27)	(0.83)	(1.38)	(1.32)	(0.89)	(1.42)	(1.23)	(0.93)	(1.28)	(1.28)	(0.86)	(1.27)	(1.26)	(0.89)	(1.16)
Young	0.84	0.78	0.08	0.71	0.86	0.06	-	-	-	0.67	0.74	0.17	0.72	0.75	0.15
Old	0.62	0.88	0.07	0.60	0.89	0.06	-	-	-	0.63	0.78	0.14	0.53	0.86	0.13
$\frac{\text{Young}}{\text{Old}}$ (Ratio)	(1.36)	(0.89)	(1.17)	(1.19)	(0.97)	(1.05)	-	-	-	(1.06)	(0.95)	(1.22)	(1.35)	(0.88)	(1.18)

when introducing ML, and calibration and some bias mitigation method (such as equalized odds in this study) needs to be used.

ACKNOWLEDGMENTS

This work has been partially supported by the HUMAINT programme (Human Behaviour and Machine Intelligence), Centre for Advanced Studies, Joint Research Centre, and European Commission. The project leading to these results has received funding from "la Caixa" Foundation (ID 100010434), under the agreement LCF/PR/PR16/51110009.

REFERENCES

[1] Antonio Andrés-Pueyo, Karin Arbach-Lucioni, and Santiago Redondo. 2018. The RisCanvi: a new tool for assessing risk for violence in prison and recidivism. *Recidivism Risk Assessment: A Handbook for Practitioners* (2018), 255–268.
[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.
[3] Richard Berk. 2012. *Criminal justice forecasts of risk: A machine learning approach.* Springer Science & Business Media.
[4] Richard Berk. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *J. of Experimental Criminology* 13, 2 (2017), 193–216.
[5] Richard Berk. 2019. Accuracy and fairness for juvenile justice risk assessments. *J. of Empirical Legal Studies* 16, 1 (2019), 175–194.
[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.
[7] Richard Berk and Jordan Hyatt. 2015. Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter* 27, 4 (2015), 222–228.
[8] Richard A Berk, Susan B Sorenson, and Geoffrey Barnes. 2016. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *J. of Empirical Legal Studies* 13, 1 (2016), 94–115.
[9] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36, 1 (2009), 21–40.
[10] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[11] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).
[12] S. Corbett-Davies and S. Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).
[13] K.P. Dahle, J. Biedermann, R.J. Lehmann, and F. Gallasch-Nemitz. 2014. The development of the Crime Scene Behavior Risk measure for sexual offense recidivism. *Law and human behavior* 38, 6 (2014), 569.
[14] Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. 2018. The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky. *Available at SSRN 3168452* (2018).
[15] S.L. Desmarais, K.L. Johnson, and J.P. Singh. 2016. Performance of recidivism risk assessment instruments in US correctional settings. *Psychological Services* 13, 3 (2016), 206.
[16] Sarah Desmarais and Jay Singh. 2013. Risk assessment instruments validated and implemented in correctional settings in the United States. (2013).
[17] W. Dieterich, C. Mendoza, and T. Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc* (2016).
[18] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. 2000. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment* 12, 1 (2000), 19.
[19] Melissa Hamilton. 2019. The sexist algorithm. *Behavioral sciences & the law* 37, 2 (2019), 145–157.
[20] M. Hardt, E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
[21] Philip D Howard and Louise Dixon. 2012. The construction and validation of the OASys Violence Predictor: Advancing violence risk assessment in the English and Welsh correctional services. *Criminal Justice and Behavior* 39, 3 (2012), 287–307.
[22] Danielle Leah Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. (2017).
[23] J. Kleinberg, S. Mullainathan, and M. Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[24] Carolin Kröner, Cornelis Stadtland, Matthias Eidt, and Norbert Nedopil. 2007. The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. *Criminal Behaviour and Mental Health* 17, 2 (2007), 89–100.
[25] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).
[26] Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo. 2020. Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law* (2020), 1–37.
[27] Arvind Narayanan. 2018. 21 fairness definitions and their politics. *presenterad på konferens om Fairness, Accountability, and Transparency* 23 (2018).
[28] Rachael T Perrault, Gina M Vincent, and Laura S Guy. 2017. Are risk assessments racially biased?: Field study of the SAVRY and YLS/CMI in probation. *Psychological assessment* 29, 6 (2017), 664.
[29] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.
[30] M. Rettenberger, M. Mönichweger, E. Buchelle, F. Schilling, and R. Eher. 2010. The development of a screening scale for the prediction of violent offender recidivism. *Monatsschrift für Kriminologie und Strafrechtsreform* 93, 5 (2010), 346–360.
[31] J.P. Singh, D.G. Kroner, J.S. Wormith, S.L. Desmarais, and Z. Hamilton. 2018. *Handbook of recidivism risk/needs assessment tools.* John Wiley & Sons.
[32] Jay P Singh, Sarah L Desmarais, Cristina Hurducas, Karin Arbach-Lucioni, Carolina Condemarin, Kimberlie Dean, Michael Doyle, Jorge O Folino, Verónica Godoy-Cervera, Martin Grann, et al. 2014. International perspectives on the practical application of violence risk assessment: A global survey of 44 countries. *Int. J. of Forensic Mental Health* 13, 3 (2014), 193–206.
[33] Jennifer L Skeem and Christopher T Lowenkamp. 2016. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology* 54, 4 (2016), 680–712.
[34] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. [n.d.]. Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. In *Proc. of ICAIL'19*.
[35] Jeffrey Todd Ulmer and Darrell J Steffensmeier. 2014. The age and crime relationship: Social variation, social explanations. In *The nurture versus biosocial debate in criminology: On the origins of criminal behavior and criminality.* SAGE Publications Inc., 377–396.
[36] B. Woodworth, S. Gunasekar, M.I. Ohannessian, and N. Srebro. 2017. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081* (2017).
[37] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proc. of the 26th Int. Conf. on WWW*. 1171–1180.