Efficiency and Fairness in Recurring Data-Driven Risk Assessments of Violent Recidivism

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

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

ABSTRACT

In this paper, we consider the prediction of violent recidivism in criminal justice as currently done through machine learning methods. Specifically, we consider sequential evaluations performed on jail inmates with a state-of-the-art risk assessment instrument, RisCanvi. In this protocol, evaluations are done periodically every six months to all inmates. We study a scenario in which the inter-evaluation period depends on the characteristics of each inmate. In this scenario, only a fraction of the inmates, those with the highest probability of having changed risk, are selected for the next evaluation. Our work is based on a cost-benefit analysis which leads to fewer evaluations in exchange for some missed/undetected changes. When modeling risk change, we obtain prediction models with AUC in the order of 0.74-0.78, which can be used to schedule evaluations leading to a small number of missed changes (about 14%) by performing half of the evaluations (50%). This allows freeing resources and staff for other tasks. Importantly, we analyze if this method leads to discriminatory outcomes across some characteristics, including disparate impact in the evaluation rates along nationality and age. By adjusting decision boundaries we are able to mitigate the disparate impact and ensure equality in the rate of evaluation. Even after mitigation, missed changes remain small (about 15%) while still halving the number of evaluations needed.

CCS CONCEPTS

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

KEYWORDS

fairness, machine learning, violent risk assessment

ACM Reference Format:
Marzieh Karimi-Haghighi and Carlos Castillo. 2021. Efficiency and Fairness in Recurring Data-Driven Risk Assessments of Violent Recidivism. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21), March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3412841.3441975

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 [29]. These tools are used by probation and parole officers, police, and psychologists to assess the risk of harm, sexual, criminal, and violent offending in more than 44 countries [39]. The main purpose of violence risk assessment tools is to prevent criminal violence and its consequences, but they also help prison management identify offenders with a greater risk of recidivism and allocate rehabilitation efforts accordingly. Ideally, accurate risk assessment may help place low-risk defendants in alternative programs to prison [1].

In comparison to traditional prediction methods and unstructured clinical judgments, risk assessment tools offer superior accuracy and performance [22]. 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 [4]. Such advances have effectively increased the use of tools based on Machine Learning (ML) in criminal justice decisions for risk forecasting [5, 8, 9]. ML-based systems provide automatic methods that can improve accuracy and efficiency by discovering and exploiting regularities in historical (training) data [32].

Today, various semi-structured protocols for assessing risk of recidivism can be found in different countries including the U.S. [18], the U.K. [25], Canada [31], Austria [37], and Germany [16]. 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. In the RisCanvi protocol, each inmate is evaluated every six months and each evaluation results in four scores predicting four outcomes: (i) violent recidivism, (ii) self-directed harm/violence, (iii) violence within the prison facilities, and (iv) breaking of prison permits [1].

Our contribution. Performing risk evaluations for all of the inmates every six months is an expensive and time-consuming task. We observe that the risk score for "Violent Recidivism" (hereinafter referred to as REVI) changes differently over time depending on the initial risk. Therefore, we study a scenario in which a decision on the next evaluation for each inmate is taken using an ML-based prediction of the risk change. To this purpose, three ML models are generated for the prediction of change within 6, 12, and 18 months. The ML models are created using risk factors (details in Section 3), the current risk score (REVI) generated by the RisCanvi protocol using those risk factors, and demographic factors.

SAC '21, March 22–26, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8104-8/21/03…$15.00
https://doi.org/10.1145/3412841.3441975

We perform a simulation in which only those inmates with the highest probabilities of risk change are selected for the next assessment. We show that this can halve the number of evaluations with a relatively small number of missed/undetected risk changes.

We also perform an evaluation of potential algorithmic bias introduced by this method. Given that ML-based models may lead to unfairness [14, 15, 41], we compare the impact of our data-driven scheduling of risk assessment along nationality and age. This impact is investigated along four metrics: accuracy differences, inequality in the missed changes which can be considered analogous to a notion of disparate mistreatment [44], and disparate impact in the rate of evaluations or fraction of unnecessary evaluations. We address these disparities through an algorithmic discrimination mitigation procedure, which equalizes evaluation rates across nationality and age. As expected, in exchange of evaluations rate parity, there is an increase in missed changes. As we will show, this increase is small.

We remark that there are many similar domains in which professionals need to perform periodic appraisals, potentially with the assistance of an algorithm, including education, public health, allocation of social benefits, and information security. In all cases where recurring data-driven risk assessment is used to make these kinds of decisions for individuals, the frequency of these assessments is key to achieve a balance of costs and benefits, and it is important to consider and mitigate the potential algorithmic bias that may be inadvertently introduced when seeking to reduce such frequency.

The rest of this paper is organized as follows. In Section 2, a brief description of the related work is presented. In Section 3, some explanations regarding the RisCanvi risk assessment tool, the data set used in this study, and violent recidivism are provided. The methodology including the model-level evaluation, system-level evaluation, and fairness evaluation are presented in Section 4. The results related to each of the evaluation metrics are given in Section 5. To mitigate the discrimination, a procedure is suggested in Section 6. Finally, the obtained results are discussed in Section 7 and the paper is concluded in Section 8.

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.

In the US, a widely-used program named Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) has been found to have biases across races and genders. In seminal research done by investigative journalism organization ProPublica [2, 33] it was concluded that the COMPAS risk assessment tool is biased against African American defendants. A follow-up study [23] analyzed the fairness of COMPAS in terms of predictive parity, and 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 [20]. 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) do not exhibit racial bias in the recidivism prediction [40]. Similarly, in risk assessment tools used in juvenile probation decisions, such as the Structured Assessment of Violence Risk in Youth (SAVRY) and the Youth Level of Service/Case Management Inventory (YLS/CMI), no significant racial bias has been found in the prediction of re-offending, except for a higher score in African American youth compared to White youth in the YLS/CMI scale related to official juvenile history [35]. In a more recent study focused on SAVRY [41], it is shown that although ML models could be more accurate than the simple summation used to compute SAVRY scores, they would introduce discrimination against some groups of defendants compared to the current method.

In general, it is impossible to maximize fairness and accuracy at the same time [6, 7]. There are many different definitions of algorithmic fairness [34], some of which are incompatible with one another. It is impossible to satisfy all of them simultaneously, hence, there are necessary trade-offs between different metrics [7, 13, 30]. In this regard, some studies [24, 43, 44] try to mitigate potential algorithmic discrimination by satisfying equalized odds or in other words avoiding disparate mistreatment along different sensitive groups. In the methodology used by Zafar et al. disparate treatment can also be avoided simultaneously with disparate mistreatment since sensitive feature information is not used while making decisions, which make it more applicable for the scenarios when the sensitive attribute information is not available. Also, several studies [26–28, 45] tried to approach statistical parity in which the same probability of receiving a positive-class prediction is considered for different groups. In addition, due to the importance of the calibration in risk assessment tools [7, 20], some previous work has also tried to minimize error disparity across groups while maintaining calibrated probability estimates [36].

As explained, our work is based on a cost-benefit analysis which results in fewer evaluations in exchange for some missed (undetected) changes. Thus, to mitigate potential algorithmic bias there is a trade-off between some fairness metrics; mitigating disparate impact in the evaluation rates and disparate mistreatment regarding undetected risk changes along groups. Since simultaneous satisfaction of both measures is impossible we try to mitigate the disparate impact in the rate of evaluation across groups while keeping the fraction of missed changes small.

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]. It was designed jointly by professionals working in the prison system, including lawyers, social workers, criminologists, and psychologists, similarly to other risk assessment tools [10, 11]. RisCanvi is not a questionnaire. Instead, each inmate is interviewed by professionals, who are responsible for analyzing different areas of the inmate's progress through the lens of some risk factors. Each evaluation requires multiple interviews by several professionals spread along several days. RisCanvi interviews are coded by trained professionals and a system generates a risk score; a committee accepts or modifies this score and decides the next action, intervention, or program.

In the 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. To determine the probability of each outcome, a unique predictive algorithm was designed. Each predictive algorithm incorporates various risk factors and three additional variables: age, gender, and country of origin (Spanish or foreign).

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 for both versions are shown on Table 1.

Risk items are grouped into five different categories: Criminal/Penitentiary, Biographical, Family/Social, Clinical, and Attitudes/Personality. These items can also be divided into static factors (such as "criminal history in their family" and "age at first violent offense") and dynamic factors (such as "member of socially vulnerable groups" and "pro-criminal or antisocial attitudes"). In the screening version RisCanvi-S, some risk items are directly taken from RisCanvi-C and others are a combination of items [1].

RisCanvi 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. Generally, the screening version RisCanvi-S is applied to all inmates when they enter the prison. The outcome of RisCanvi-S can be "high-risk" or "low-risk." If the outcome is low-risk for all four criteria, the same RisCanvi-S protocol is repeated after six months. Otherwise, in the case of high-risk levels or significant change in an inmate's situation, the complete version RisCanvi-C is applied. The outcome of RisCanvi-C can be "high-risk," "medium-risk," or "low-risk." When the risk levels measured by RisCanvi-C are medium or high, the next evaluation is again a RisCanvi-C; otherwise, the RisCanvi-S is used [1].

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. We kept only offenders for which nationality information was recorded, that comprises 2,634 offenders. Among this population, 256 inmates had violent recidivism after being released. The data includes all of the information for the two RisCanvi versions (RisCanvi-S and RisCanvi-C). All inmates were evaluated at least once, and depending on the time they spend in prison, 46% had a second evaluation, 18% a third one, and less than 6% had four to eight evaluations. On average, inmates with only one evaluation remain for about three months in prison, while inmates with four evaluations on average spend two years before being released on parole or regaining freedom. There is no evaluation after an inmate's release.

Handling missing items. In the RisCanvi data, there were some missing items. Static items were replaced with the value of their counterparts from the previous or next evaluations. These static items were 7 items from the 43 RisCanvi-C risk factors in Table 1 (items 8, 16, 22, 23, 32, 33, and 39). Moreover, items with a yes/no/uncertain answer in which there was a missing item, had missing values replaced with "uncertain." After the above replacements, we removed the cases with irreplaceable missing items from the sample, leaving 2,582 people in the final data set.

Table 1: RisCanvi Risk Factors, with Items Related to Violent Recidivism Marked in Boldface

	RisCanvi Complete items (S = shared with Riscanvi Screening)
(1)	**Violent base offense**
(2)	Age at the time of the base offense
(3)	**Intoxication during performing the base offense**
(4)	Victims with injuries
(5)	Length of criminal convictions
(6)	Time served in prison
(7)	**History of violence** (S)
(8)	Start of the criminal or violent activity (S)
(9)	**Increase in frequency, severity and diversity of crimes**
(10)	**Conflict with other inmates**
(11)	Failure to accomplishment of penal measures
(12)	**Disciplinary reports**
(13)	Escape or absconding
(14)	Grade regression
(15)	Breaching prison permit
(16)	**Poor childhood adjustment**
(17)	**Distance from residence to prison**
(18)	**Educational level**
(19)	Problems related with employment
(20)	**Lack of financial resources** (S)
(21)	**Lack of viable plans for the future**
(22)	**Criminal history of family or parents**
(23)	Difficulties in the socialization or development in the origins family
(24)	Lack of family or social support (S)
(25)	Criminal or antisocial friends
(26)	Member of social vulnerable groups
(27)	**Relevant criminal role**
(28)	**Gender violence victims (only women)**
(29)	Responsibility for the care of family
(30)	**Drug abuse or dependence**
(31)	**Alcohol abuse or dependence**
(32)	Severe mental disorder
(33)	Sexual promiscuity and/or paraphilia
(34)	**Limited response to psychological and/or psychiatric treatments** (S)
(35)	Personality disorder related to anger or violent behaviour
(36)	Poor stress coping
(37)	**Self-injury attempts or behaviour** (S)
(38)	**Pro criminal or antisocial attitudes**
(39)	**Low mental ability**
(40)	**Recklessness**
(41)	Impulsiveness and emotional instability
(42)	**Hostility**
(43)	**Irresponsibility**
	Other RisCanvi Screening items
(1)	Institutional/prison misconduct
(2)	Escapes or breaches of permits
(3)	Problems with drugs or alcohol use
(4)	Hostile or pro criminal attitudes

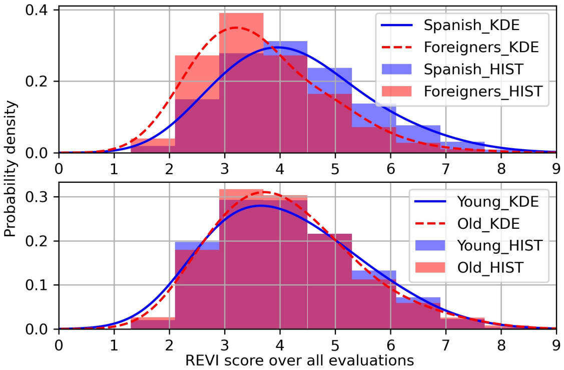

Figure 1: Violent recidivism score (REVI) distribution by nationality and age. Foreigners tend to have slightly lower REVI scores than nationals. Both "young" (≤ 30 years old) and "old" (> 30 years old) have similar REVI scores. Smooth curves are obtained by Kernel Density Estimation (KDE).

Table 2: Violent Recidivism Rate (Average)

Spanish	12.4%	"Young" (age ≤ 30)	12.7%
Foreigners	8.9%	"Old" (age > 30)	10.8%

3.3 Violent Recidivism (REVI)

Violent crimes are more costly to victims and the criminal justice system compared to other crimes [38]. Also, violent recidivism can be more clearly established and hence the ground truth is more reliable. Therefore, in this work we focus on RisCanvi to assess Violent Recidivism ("REVI" in the protocol) risk in sentenced inmates. REVI risk is an outcome predicted using a sub-set of risk factors shown in boldface on Table 1 (23 out of the 43 risk factors of the RisCanvi-C version), plus two demographic features (gender and nationality). In RisCanvi-C, the REVI score has been computed by applying the summation of these features in a hand-crafted formula, then using two cut-offs, obtaining three REVI risk levels (details in [1]).

First, we compare the distribution of REVI risk scores by nationality and age groups. We do not consider a grouping by gender as the number of women in our sample is too small to draw robust conclusions. Fig. 1 shows the distribution of REVI risk scores per group, while the average recidivism rate per group is shown on Table 2. For age groups we use 30 years old 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., [42]). This age is also used as a cut-off for young and old people in the design of the RisCanvi protocol. In our dataset, the majority of the population are Spanish nationals (68%) and older than 30 years old (67%). As can be seen in Fig. 1, foreigners tend to have lower REVI risk scores compared to Spanish. Also, the distribution of REVI score by age shows that old and young inmates have similar risk score distributions. In this dataset, we observe that foreigners, which have lower risk, have less tendency to change in REVI risk compared to Spanish nationals. For the same reason, inmates older than 30 years old are slightly less likely to change in REVI risk compared to younger offenders.

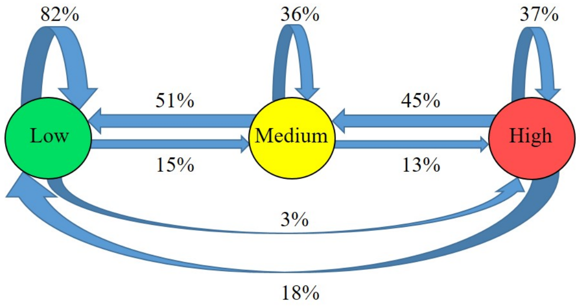

Figure 2: REVI variation in 12 months. Low-risk inmates tend to have the same risk in successive evaluations, whereas medium- and high-risk inmates tend to exhibit less risk.

Second, given our goal is to study the consequences of selectively re-evaluating to reduce the overall number of evaluations, we look at how risk changes. Fig. 2 depicts REVI risk changes in RisCanvi evaluations separated by 12 months intervals. We also obtained the REVI variations in 6 and 18 months intervals (figure omitted for brevity). In general, we note a larger probability of REVI risk changes when the interval is longer. Also, when the risk changes, there is more tendency to decrease. For medium- and high-risk inmates we observe a tendency to lower risk levels in the next evaluation, and for low-risk inmates a tendency to continue being evaluated as low-risk. This can be due to the effects of the rehabilitation and other interventions done while in the prison and it goes contrary to the incarceration effects noted in some works [21].

Third, to have more insights on the RisCanvi dataset and REVI risk scores, we create a new machine learning classifier using 43 risk factors, three demographic features (gender, age, and nationality), and REVI risk level (low, medium, and high). We use an off-the-shelf multi-layer perceptron as learning scheme, which performed better than other methods we tested for this task (including logistic regression and support vector machines). The cases considered in this model are 2,028 (out of 2,582) who are sentenced (not awaiting trial), that were released at most 9 months after their last RisCanvi evaluation, and for which violent recidivism (or its absence) was recorded at most two years after their freedom. Using 5-fold cross validation, the average AUC of the model is 0.69. In comparison, RisCanvi-C obtains an AUC of 0.68. These values are in line with that of similar tools used in other countries, which tend to have AUC values in the range of 0.57-0.74 [12, 17, 19].

4 METHODOLOGY

Normally, each inmate is evaluated every six months; we test the effects of performing less RisCanvi evaluations by selectively postponing the evaluation of an inmate for two periods or three periods (i.e., 12 months or 18 months). As ground truth, the cases over which we test are only inmates who actually received four evaluations regularly in an 18 month period, so we know whether their risk changed or not.

Three ML models corresponding to periods of 6, 12, and 18 months, are created to predict the necessity for a new evaluation at the end of a period. We use different ML methods, such as logistic regression, multi-layer perceptron (MLP), and support vector machines (SVM). The features used for the time prediction models are Violent Recidivism (REVI) items (boldface in Table 1), REVI score, gender, nationality, and age at the time of evaluation. Additionally, in 12- and 18-month models the output(s) from the shorter-period model(s) are used as additional features.

The whole data is split into two sets. The first set is divided into training and validation and used to create the 6, 12, and 18 months risk change models and for performing model-based evaluation. The second set is used for both model- and system-level evaluations. In the system-level evaluation, this set is used to schedule the evaluation of inmates using the prediction models as explained next. In the simulation, every six months, a fraction σ (the selection rate) of the inmates with the highest probability of REVI risk change (obtained in the previous six months period using ML models) are selected for evaluation. Those evaluated have their REVI risk change probabilities recomputed for the next six months.

The split for model-level and system-level evaluation is done k times using k-fold cross-validation, reporting average results. The part for model-based evaluation is also split using k-fold cross validation.

4.1 Model-Level Evaluation Methodology

We consider changes in REVI risk level between two evaluations separated by a time interval (6, 12, or 18 months). This is modeled as a binary classification task in which we have to predict whether there will be a change or not at the end of the period. If risk changes we have a positive example, if risk does not change we have a negative example. The predictive performance of the ML models is evaluated using Area Under the ROC curve (AUC-ROC).

4.2 System-Level Evaluation Methodology

System-level evaluation is done through a simulation of 18 months. In the simulation, at time 0, the three ML models (6-, 12- and 18-month models) are applied to the second set (introduced in Section 4) and three series of predictions for the next 6, 12, and 18 months are obtained for each inmate. Then in each six months period, a fraction σ of the inmates with the highest probabilities of REVI risk change are selected for evaluation and the rest have their evaluation postponed. Whenever selected individuals are evaluated, we apply the ML models over their actual RisCanvi evaluation (which is known), and based on the new obtained predictions, the old predictions are updated.

By selecting only a part of inmates each time, there will be some omissions or *missed changes*: cases who experience REVI risk change but are not evaluated and hence not detected. As risk change leads to a positive class, *Missed changes* can be interpreted as False Negative Rate (FNR) and formulated as *undetected changes* divided by *total changes*.

Thus, we undertake a cost-benefit analysis. The *cost* is the fraction of inmates who experience an undetected risk change. The *benefit* (equal to $1 - \sigma$) is the fraction of evaluations that are not done, i.e., the resources saved because not all inmates have to be evaluated. The cost of the baseline (current method) is 0, as all risk changes are detected, and its benefit is also 0, as this is equivalent to have a selection rate $\sigma = 1$.

We compute the *cost* (*missed changes*) in two ways: cases with undetected REVI risk increase and cases with undetected REVI risk decrease. Studying missed risk increases is important since the outcome can be dangerous to society. Also, postponing the evaluation of inmates who have less risk now may have negative psychological effects on the inmates, and can have a negative impact on their rehabilitation. Furthermore, to study people with REVI risk increase and decrease more precisely, we create ML models for each group separately.

An additional metric we compute is the *average number of evaluations* per inmate, a figure that we compute globally as well as per-group as explained next. This is a number between 1 (inmate is evaluated at the beginning and at some point in the next 18 months) and 3 (inmate is evaluated at the beginning, and then at 6, 12, and 18 months). Note we do not count the initial evaluation in this computation because it is shared among all settings.

Finally, we also compute the *average number of unnecessary evaluations*, which are REVI risk evaluations in which the outcome is the same risk level as the previous evaluation. Only a perfect predictive model (an oracle) could reduce this number to zero.

4.3 Algorithmic Fairness Evaluation Methodology

Finally, we consider *algorithmic fairness* by comparing metrics across groups. First, we study whether ML models show any discrimination in the prediction of REVI change against "Spanish" or "foreigners" and "young" or "old" inmates. Second, we check the disparate impact in the average number of evaluations along nationality and age. Third, for the obtained rate of undetected changes, we study if there is a disparate mistreatment (FNR discrepancy) between nationality and age sub-groups. Finally, we study if there is a disparate effect in terms of the average number of unnecessary evaluations between these groups.

5 RESULTS

5.1 Model-Level Evaluation

To evaluate the predictive performance of the ML models, the validation data of the first set and the whole second set (introduced in Section 4) are used. Among MLP (Multi-Layer Perceptron), logistic regression and support vector machines, the best results in terms of the AUC-ROC were obtained using MLP with a single hidden layer having 100 neurons. Hence, the non-MLP based models are omitted for brevity. In Table 3, the results in terms of the AUC-ROC are presented. According to the AUC values, we can see that the three ML models (6-,12-, and 18-month) have good accuracy. We remark that AUC values are dominated by low-risk individuals, who are the majority in this data (the average percentage of low-risk people is 70%).

Table 3: AUC of Risk Change Prediction Models

6-month model	12-month model	18-month model
0.78	0.75	0.74

Table 4: AUC of Risk Change Prediction Models per Group

Model	6-month	12-month	18-month
Spanish	0.75	0.70	0.70
Foreigners	0.85	0.85	0.78
Young (age ≤ 30)	0.77	0.74	0.76
Old (age > 30)	0.79	0.75	0.73

5.2 System-Level Evaluation

5.2.1 Missed/Undetected Changes. As mentioned, given only a fraction σ of inmates is evaluated, there are missed or undetected changes. In Fig. 3, the fraction of REVI missed changes (increase or decrease), REVI missed increases and REVI missed decreases are shown for different selection rates of the inmates. We see a much smaller number of missed changes compared to selecting inmates at random (the result for random selection of the inmate is shown by "Chance" curve in Fig. 3). Also, the missed values for REVI change, increase, and decrease are very similar. This curve represents a series of trade-offs, and the specific trade-off should be chosen by the experts depending on the cost they assign to different aspects. We consider a selection rate of 50% in the following for concreteness, but remark that other selection rates could be chosen and would be analyzed in the same manner. Thus, by selecting 50% of the inmates with the highest probability of REVI change each time, we would miss about 12% to 15% of changes. Moreover, we note that we are more accurate at avoiding missed changes in a short time frame (6 months) compared to longer periods (12 or 18 months) and by selecting more than 50% of the inmates in 6-month model, there would be zero missed changes, because REVI tends to change in longer time intervals, as explained in Section 3.3.

In addition, by evaluating the ML models created separately for the two groups with REVI risk increase and decrease (figures and details omitted for brevity), we conclude that the REVI risk decrease model shows more accuracy and less missed values compared to REVI change and REVI increase models.

5.2.2 Evaluations Per Inmate. Our goal is to reduce the average number of evaluations performed for each inmate. According to our results in Fig. 4, the average number of evaluations performed by our method is smaller than the 3 evaluations required by standard RisCanvi in an 18 months period. For instance, by selecting $\sigma \approx 50\%$ of inmates (those with the highest probability of REVI change), there would be 1.5 evaluations per inmate on average.

Unnecessary evaluations are situations where an evaluation is performed and yields the same risk score as the previous evaluation. Our models lead to less unnecessary evaluations on average compared to the RisCanvi (figures omitted for brevity). Again, by selecting $\sigma \approx 50\%$ of the inmates for evaluation, the average number of unnecessary evaluations per inmate would be close to 1.0, which is less than standard RisCanvi (2.4 unnecessary evaluations per inmate on average).

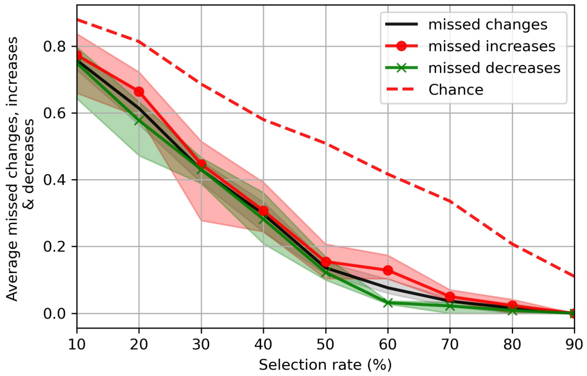

Figure 3: REVI missed changes. There is a much smaller number of missed changes compared to selecting inmates at random ("Chance").

5.3 Algorithmic Fairness Evaluation

In Table 4, the results for the analysis of equity in accuracy (AUC) are shown for nationality (Spanish and foreigners), and age (young and old inmates) groups. The AUC results of the ML models show more accuracy for foreigners than for Spanish nationals in general, despite the latter comprising about 68% of this sample. For the age groups, the difference is small.

Next we check if there is parity in the average number of evaluations per Spanish and foreigner in Fig. 4, for various selection rates. We observe that on average Spanish nationals and foreigners receive 1.69 (with a **spread** of 0.12 between the min and max among folds) and 1.08 (spread: 0.21) evaluations respectively for the selection rate of 50%. These results show more average number of evaluations per Spanish compared to foreigners. The same analysis for "young" vs "old" inmates is shown in Fig. 4. The results show that for the selection rate of 50%, there are 2.08 (spread: 0.18) and 1.20 (spread: 0.08) average number of evaluations per young and old respectively which represent more average number of evaluations per young than old inmate.

According to the results obtained for the average number of unnecessary evaluations, for the selection rate of 50%, on average Spanish with the value of 1.07 have more unnecessary evaluations than foreigners with the value of 0.7, since they have more average number of evaluations (Fig. 4). Also considering a selection rate of 50%, the results for "young" and "old" inmates are 1.37 and 0.75 respectively, which shows that on average there are more unnecessary evaluations for younger inmates; this is consistent with the results of the average number of evaluations in these sub-groups (Fig. 4).

Furthermore, the missed changes (FNR) for each sub-group of nationality and age are shown in Fig. 5. For the selection rate of

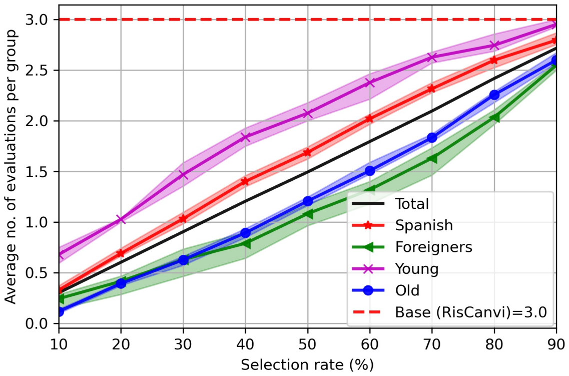

Figure 4: Average number of evaluations per person. Our method leads to a smaller number compared to the standard RisCanvi which requires 3 evaluations in an 18 months period. However, without mitigation measures for algorithmic bias, the evaluation rate is different across groups.

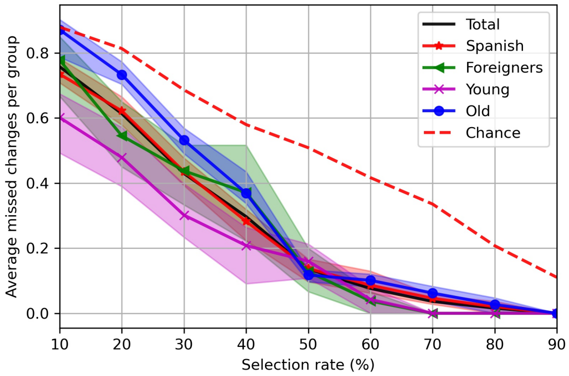

Figure 5: Average missed changes per group. For the selection rate of 50%, the missed changes difference in nationality (Spanish vs foreigners) and age (young vs old) is too small.

σ =50% missed changes for Spanish and foreigners are 0.14 (spread: 0.07) and 0.13 (spread: 0.13) respectively. For young and old subgroups, the results show 0.16 (spread: 0.10) and 0.12 (spread: 0.04) missed changes respectively. For this particular cut-off value, and in general for selection rates larger than 50%, differences in missed changes are relatively small.

6 MITIGATING ALGORITHMIC BIAS

The method we have described could introduce a disadvantage for a group of inmates if that group is consistently evaluated more often or less often than another. We would prefer to select inmates for evaluation at the same rate σ independently of their nationality, age, or other characteristics. In our experiments, by moving the decision boundary we select inmates so that the selection rate is similar for different nationality and age groups. First, we select a fraction σ of inmates having the highest probability of Violent Recidivism (REVI) change from both groups by nationality (nationals and foreigners). Second, we add cases with high probability of REVI change and remove cases with low probability of REVI change until both age groups ("young" and "old") are equalized.

The rate of missed changes after applying the mitigation process increases by about three percentage points (figures omitted for brevity). By selecting $\sigma \approx 50\%$ of the inmates with the highest probability of REVI change, we would miss between 14% to 18% in REVI changes, increases, and decreases, compared to 12%-15% missed changes before bias mitigation. The obtained results for the selection rate of 50% show missed changes of 0.16 and 0.13 for Spanish and foreigners respectively which represents a small range difference. These results for young and old inmates are 0.22 and 0.10 respectively which shows more missed changes for younger inmates.

The average number of evaluations per inmate after applying the bias mitigation procedure increases in the case of foreigners, and decreases in the case of Spanish nationals. This also decreases for young inmates and increases for old inmates. This is because we are correcting a disparity that was present before applying the mitigation. Our results show that for the selection rate of 50% of the inmates with the highest probability of REVI change, on average there are about 1.7 evaluations per inmate (same value for young and old inmates, 1.6 for Spanish nationals and 1.9 for foreigners); compare this to 1.5 evaluations per inmate before bias mitigation.

The average number of unnecessary evaluations per person after bias mitigation is 1.2 (1.0 for young inmates, 1.2 for old inmates, 1.0 for Spanish nationals and 1.5 for foreigners) for the selection rate of 50%; compare this to about 1.0 unnecessary evaluations per inmate before bias mitigation. The reason is that balancing the evaluation rate caused less evaluations and accordingly less unnecessary evaluations for Spanish nationals who were evaluated more often in the scenario without bias mitigation. Something similar happens in the case of the unnecessary evaluations of "young" vs "old" inmates: the values are lower for young inmates and higher for old inmates.

Finally, if we wanted to ensure a specific bound on the number of missed changes, this would require a particular minimum selection rate. Fig. 6, shows the evaluation rate needed to have on expectation a certain amount of missed changes before and after the mitigation. According to the results, missed change differences are small before and after the mitigation, and in particular for the selection rate of 50%, the difference is almost zero.

7 DISCUSSION

We used ML models to predict changes of violent recidivism risk, these models have AUC in the range of 0.74-0.78. In the cost-benefit analysis of selecting the inmates for the next RisCanvi evaluation, we obtained a cost, in terms of missed changes, of nearly 14% when selecting the top σ =50% of the inmates with the highest probability of Violent Recidivism (REVI) change. The benefit is that the number of evaluations is halved. Other points in the cost-benefit trade-off curve can be used. Marginal benefits (further drops in missed changes) are decreasing, showing some saturation effect after reaching about σ =70% selection rate.

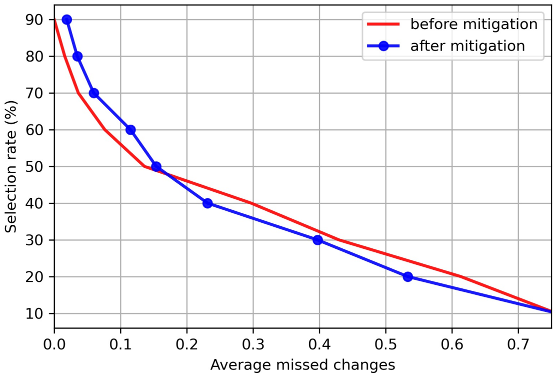

Figure 6: Evaluation rates per missed change before and after the mitigation. For the selection rates more than 50%, less than 5% more evaluations are needed to have no variation in the missed changes after the mitigation.

We observe this method introduces some differential treatment across groups such as disparate impact in the evaluation rates and disparate mistreatment with regard to undetected risk changes (false negative rates). Specifically, as the results showed in Fig. 4, the average number of evaluations that a Spanish national must undergo is more than a foreigner. The source of this difference is that according to the results obtained in Section 3.3, foreigners are less likely to change in REVI risk, so they should expect to be less selected for the next evaluation. Similarly, the difference in the average number of evaluations along age (Fig. 4), can be traced to the same reasons, a lower tendency in old offenders to change in REVI risk.

Since there is a trade-off in mitigating both disparate impact in evaluation rate and disparate mistreatment in missed changes simultaneously, by moving the decision boundary, we mitigated the disparity in the evaluation rates along both nationality and age groups with a small additional loss of missed changes.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we employ ML-based methods to select the inmates for the next evaluation of the Violent Recidivism (REVI) risk in the RisCanvi protocol. These models showed good results in terms of AUC (0.74-0.78), which resulted in fewer evaluations per inmate compared to the standard RisCanvi, which in turn leads to save time, expenses and staff in the evaluations. This benefit has been obtained in exchange for some missed changes (about 14% when selecting 50% of the inmates with the highest probability of REVI change).

Furthermore, analyzing the fairness of the ML models along nationality (Spanish and foreigners) and age (young and old) led to the following results: in terms of AUC, the models are more accurate for foreigners than Spanish nationals and there is no significant difference in age sub-groups. In terms of missed changes (false negative rates), for the selection rate of 50%, the disparate mistreatment is less than 0.04 among both nationality and age sub-groups. There is a disparate impact in the average number of evaluations which shows lower number of evaluations in foreigners and older inmates on average. This also translates to a difference in the average number of unnecessary evaluations per group.

Applying a mitigation method to gain parity in the rate of evaluations along nationality and age leads to a small increase in missed changes which is less than one percentage point for the selection rate of 50%. We obtained parity in the average number of evaluations per inmate along both nationality and age which is 1.7; about half of the evaluations done in RisCanvi.

The method used in this study can also be used for other RisCanvi criteria: self-directed violence, violence to other inmates or prison staff, and breaking prison permits to see if there is still such a possibility to perform less evaluations in exchange for a small number of missed changes, while preserving equality between different groups. We must note, however, that our work is validated on data from inmates that have four evaluations and spend on average two years (or more) in prison, and might not be applicable for people receiving shorter sentences.

The freed staff time of using this method can go to programs focused on reducing the likelihood of recidivism instead of merely predicting it, something that have been advocated by researchers critical of current ML-based risk assessments [3].

The problem and approach raised in this work is general enough to be applicable in other areas where appraisals and predictions about individuals are done (e.g., education, public health, information security, immigration, social benefits, and so on).

Dataset and Reproducibility. This is a highly sensitive dataset, but access to it for research purposes is possible through a research agreement. Details will be provided in the camera-ready version. All experimental code will be made available publicly in a public code repository.

ACKNOWLEDGMENTS

This work has been partially supported by the HUMAINT programme (Human Behaviour and Machine Intelligence), Centre for Advanced Studies, Joint Research Centre, European Commission. The project leading to these results have 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] Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, and Jonathan Zittrain. 2017. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. *arXiv:1712.08238* (2017).
[4] Richard Berk. 2012. *Criminal justice forecasts of risk: A machine learning approach.* Springer Science & Business Media.
[5] Richard Berk. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology* 13, 2 (2017), 193–216.
[6] Richard Berk. 2019. Accuracy and fairness for juvenile justice risk assessments. *Journal of Empirical Legal Studies* 16, 1 (2019), 175–194.
[7] 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.

[8] Richard Berk and Jordan Hyatt. 2015. Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter* 27, 4 (2015), 222–228.

[9] Richard A Berk, Susan B Sorenson, and Geoffrey Barnes. 2016. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies* 13, 1 (2016), 94–115.

[10] Randy Borum. 2006. Manual for the structured assessment of violence risk in youth (SAVRY). (2006).

[11] Tim Brennan and William Dieterich. 2018. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). *Handbook of Recidivism Risk/Needs Assessment Tools* (2018), 49.

[12] 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.

[13] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[14] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[15] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[16] Klaus-Peter Dahle, Jürgen Biedermann, Robert JB Lehmann, and Franziska Gallasch-Nemitz. 2014. The development of the Crime Scene Behavior Risk measure for sexual offense recidivism. *Law and human behavior* 38, 6 (2014), 569.

[17] 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).

[18] Sarah Desmarais and Jay Singh. 2013. Risk assessment instruments validated and implemented in correctional settings in the United States. (2013).

[19] Sarah L Desmarais, Kiersten L Johnson, and Jay P Singh. 2016. Performance of recidivism risk assessment instruments in US correctional settings. *Psychological Services* 13, 3 (2016), 206.

[20] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc* (2016).

[21] Paul Gendreau, Claire Goggin, Francis T Cullen, and Donald A Andrews. 2000. The effects of community sanctions and incarceration on recidivism. In *Forum on corrections research*, Vol. 12. Correctional Service of Canada, 10–13.

[22] 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.

[23] Melissa Hamilton. 2019. The sexist algorithm. *Behavioral sciences & the law* 37, 2 (2019), 145–157.

[24] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[25] 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.

[26] James E Johndrow, Kristian Lum, et al. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13, 1 (2019), 189–220.

[27] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 1–6.

[28] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[29] Danielle Leah Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. (2017).

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

[31] 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.

[32] Pat Langley and Herbert A Simon. 1995. Applications of machine learning and rule induction. *Commun. ACM* 38, 11 (1995), 54–64.

[33] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).

[34] Arvind Narayanan. 2018. 21 fairness definitions and their politics. *presenterad på konferens om Fairness, Accountability, and Transparency* 23 (2018).

[35] 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.

[36] 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.

[37] Martin Rettenberger, Michael Mönichweger, Elvira Buchelle, Frank Schilling, and Reinhard Eher. 2010. Entwicklung eines Screeninginstruments zur Vorhersage der einschlägigen Rückfälligkeit von Gewaltstraftätern [The development of a screening scale for the prediction of violent offender recidivism]. *Monatsschrift für Kriminologie und Strafrechtsreform* 93, 5 (2010), 346–360.

[38] Jennifer Rubin, Federico Gallo, and Adam Coutts. 2008. Violent crime: Risk models, effective interventions and risk management. (2008).

[39] 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. *International Journal of Forensic Mental Health* 13, 3 (2014), 193–206.

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

[41] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia. (2019).

[42] 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.

[43] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081* (2017).

[44] 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 *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[45] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.