The impact of overbooking on a pre-trial risk assessment tool

Kristian Lum
kl@hrdag.org
Human Rights Data Analysis Group

Chesa Boudin
chesa.boudin@sfpd.gov
San Francisco Public Defender's Office

Megan Price
meganp@hrdag.org
Human Rights Data Analysis Group

ABSTRACT

Pre-trial risk assessment tools are used to make recommendations to judges about appropriate conditions of pre-trial supervision for people who have been arrested. Increasingly, there is concern about whether these models are operating fairly, including concerns about whether the models' input factors are fair measures of one's criminal activity. In this paper, we assess the impact of booking charges that do not result in a conviction on a popular risk assessment tool, the Arnold Public Safety Assessment. Using data from a pilot run of the tool in San Francisco, CA, we find that booking charges that do not result in a conviction (i.e. charges that are dropped or end in an acquittal) increased the recommended level of pre-trial supervision in around 27% of cases evaluated by the tool.

CCS CONCEPTS

• **Applied computing → Law**;

KEYWORDS

risk assessment, police accountability, overbooking, fairness

ACM Reference Format:
Kristian Lum, Chesa Boudin, and Megan Price. 2020. The impact of overbooking on a pre-trial risk assessment tool. In *Proceedings of FAT*2020: The ACM Conference on Fairness, Accountability, and Transparency (FAT*2020).* ACM, New York, NY, USA, 10 pages. https: //doi.org/10.1145/3351095.3372846

FAT* '20, January 27–30, 2020, Barcelona, Spain

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372846

1 INTRODUCTION

When a person is arrested a decision is made as to whether that person should be released before their case has concluded and, if so, under what conditions. Such decisions take many factors into account, including the risks of releasing that person back into the community. Actuarial risk assessment– statistical models that output a defendant's estimated risk (often risk of re-arrest or risk of failure to appear for court)– are increasingly used to aid decision-makers in this highly consequential decision.

Recently, concerns have been raised about the fairness of risk assessment models in the criminal justice context, particularly around race-based discrimination [1]. Much of the technical discussion in this area has centered around defining the fairness of a model in terms of racial parity along several, potentially conflicting, measures of predictive performance [2, 3, 5, 11, 18, 19]. Other concerns about risk assessment focus on the (un)fairness of the inputs to the models [7, 10]. For example, [9] has argued that the use of criminal history in risk assessment serves as a proxy for race and predictive models that rely upon this factor serve to justify the continued over-incarceration of minorities. [15] offers a discussion of many of the arguments supporting and refuting the fairness of risk assessment in criminal justice.

In this paper, we explore the "fairness" of one particular input of a popular risk assessment tool. Using data provided by the San Francisco Public Defender's Office (SFPD), we evaluate how often "overbooking"— booking an arrested person under charges that are higher than may be warranted by the facts of the case— influences Arnold Ventures'[1] Public Safety Assessment (PSA). Specifically, using data from a pilot run of the model in mid 2016 to mid 2017 in San Francisco, CA, we investigate how often charges that are ultimately unsubstantiated by the courts cause the PSA to make more restrictive recommendations than it would have if it had only used charges for which the person was ultimately found guilty. Allowing unsubstantiated charges to increase one's pre-trial supervision recommendations runs counter to an intuitive notion of fairness in the sense that charges that are ultimately dropped or not sufficiently supported by the facts

[1] Formerly the Laura and John Arnold Foundation

of the case to result in a conviction ought not to be used to justify more restrictive release conditions. Thus, this work assesses the fairness of the model both from the point of view of the inputs as well as their effects on the predictions and recommendations of the tool. To our knowledge, this is the first study of its kind to analyze the effect of charge-based inputs to risk assessment in this way.

In Section 2 we give context for and define overbooking for the purposes of our analysis. Section 3 gives a description of the risk assessment model we study here, the PSA, including a discussion of how booking charges are factored into the ultimate recommendations of the model. In Section 4, we describe the data we used in the analysis, and in Section 5, we describe how we analyzed the data. Section 6 presents our findings, Section 7 discusses the limitations of the the work, and Section 8 concludes.

2 OVERBOOKING

We assess the impact of overbooking on the PSA in San Francisco, CA over several months in 2016 and 2017. The local context here is important. According to a recent report based on an examination of cases taken on by the SFPD [16], "[p]eople of color receive more serious charges at the initial booking stage, reflecting decisions made by officers of the San Francisco Police Department." They concluded that the disparities in booking charges were one of the largest factors contributing to racial disparities in San Francisco's criminal justice system as a whole.

In San Francisco and elsewhere, booking officers (the police officers who formally file the initial charge) have a high degree of discretion when determining the booking charge. Specifically, when a law enforcement officer in the field makes an arrest they can either cite and release the arrestee (for certain low level-misdemeanors), or book the arrestee into county jail. When the arresting law enforcement agency transfers custody of the arrestee to the Sheriff (who runs the jail), it must provide the legal basis for the person's incarceration: the booking charge(s). Thus, typically, the officer who delivers the arrestee to the jail tells the booking deputy at the jail on what charges they are booked. There is virtually no oversight or feedback mechanism on the discretion to choose booking charges. So, for example, in a case where a person is arrested with a single illegal firearm, the arresting officer could choose to book the person on charges for three separate guns (or, in an extreme case, even murder).[2] In reviewing the case, the District Attorney's office should, in theory, discharge any inflated charges and rebook the arrestee on charges which are actually based in the evidence, but by that stage the risk assessment algorithm has already generated its recommendation using the booking charges as an input.

Many tools, like the PSA, nevertheless rely on the pre-conviction charges as an input because of the need to generate risk reports quickly. There is typically a delay of approximately 48 hours[3] between the time of booking into jail and the District Attorney's decision about whether and what files to charge (rebooking). One of the goals of incorporating risk assessment into the pre-trial process is to expedite the release of low-risk arrestees. If the algorithms could not be run until the district attorney had made a rebooking decision, everyone would have to wait in jail for days. Thus, to effectuate speedy release for some, the PSA relies on the police booking charge to generate reports within 24 hours, while the district attorney is still considering rebooking.

The ultimate objective of this analysis is to assess how often "unfair" booking charges caused the PSA to recommend excessively restrictive conditions of pre-trial supervision. To do this, we define unfair booking charges to be those charges associated with the PSA that do not go on to result in a conviction. Implicitly, this assumes that the conviction charges (the charges to which the person pleads or is found guilty) are a fair and accurate representation of the person's criminal activity. We acknowledge that this measure of the true severity of the crimes is imperfect.

On the one hand, there may exist cases where a defendant did commit the crime on which they are booked, and the defense just barely establishes reasonable doubt leading to an acquittal for those charges. It may seem disingenuous to label such charges unfair or unsubstantiated. On the other hand, as discussed in [12] and the many citations therein, "initial appraisals of dangerousness and culpability may send signals to later system actors, setting into motion a dynamic pattern of cascading disadvantage." For example, the rebooking charges are often identical to the booking charges. The rebooking charges then define the starting place from which plea deals are negotiated. Higher booking charges may follow the defendant through the whole plea process, with the plea deals they are offered being anchored to the initial charges, reducing the chances that they are offered deals with less severe charges.

Relatedly, many recent studies have found that pre-trial detention *causes* defendants to accept guilty pleas to charges that they otherwise would not have been convicted of had they been free pre-trial [6, 8, 13, 14, 20]. To the extent that the risk assessment is heeded by judges, booking charges that cause the tool to make more restrictive supervision recommendations might then indirectly be causing those charges to be "substantiated" through a guilty plea. Although conviction charges are an imperfect ground truth in that

[2] While this sounds extreme, this example is based on one of the authors' experience as a public defender.

[3] Penal Code section 825

they likely exhibit similar (though hopefully less severe) biases to the booking charges, we believe this is the best measure of appropriate charges possible given the data we have available.

3 THE PSA

The PSA developed by Arnold Ventures is a popular pre-trial risk assessment tool that is intended to "reduce the burden placed on vulnerable populations at the forefront of the criminal justice process." [4] It has been adopted by more than 40 jurisdictions around the country [17]. The purpose of the tool is to make recommendations about the appropriate level of pre-trial supervision for people who have been arrested.

In this section, we describe the PSA as it was administered during the period of our study. Since then, Arnold Ventures has modified its terminology for some of the components of the PSA as well as some of the procedures for translating the risk scores to recommendations.[4] However, it is not mandatory that jurisdictions adopt the revised version, and many jurisdictions remain using a version similar to that described below. In fact, the version of the PSA described here is still in use in San Francisco.

In order to create a risk profile for a newly arrested person and pre-trial supervision recommendation, the tool combines several separate predictions: risk of failure to appear at a future court date (FTA), risk of re-arrest for new criminal activity (NCA), and risk of re-arrest for new *violent* criminal activity (NVCA).

Each of the predictions are calculated as a function of some subset of the following:

- age at current arrest,
- whether there are pending charges at the time of the current offense,
- whether the arrested person has any prior misdemeanor convictions,
- whether the arrested person has any prior felony convictions,
- whether the arrested person has any prior convictions (misdemeanor or felony),
- the number of prior violent convictions,
- the number of prior failures to appear for court dates in the past two years,
- whether the person failed to appear prior to two years before the offense,
- whether the individual has been incarcerated as the result of a conviction in the past,
- whether the current booked offense is considered violent.

Violent charges are determined by inclusion on the extensive PSA Violent Offense list for California. We include a representative list of charges that appear on the violent offense list in Table 1 in the appendix. Both FTA and NCA predictions are conveyed on a six point scale, with higher values corresponding to a higher predicted likelihood of the corresponding undesirable outcome. The NVCA prediction is a binary prediction, with a value of one (sometimes called a NVCA flag or violence flag) corresponding to a higher predicted likelihood of future arrest for a violent crime. Each of these predictions are calculated as a linear combination of integer valued weights. This information and information on the weights associated with each factor for each type of prediction as well as extensive documentation around the implementation of the tool is available on Arnold Ventures portal[5] in the Guide to the Release Conditions Matrix [21].

Calculating the FTA, NCA, and NVCA sub-scores constitutes step (1) of a four-step process. In the subsequent three steps, these risk scores are combined with the booking charges using pre-defined sets of rules to transform the risk scores into recommendations.

In step (2) of the PSA, three determinations are made:

- whether the person was extradited for the current booked offense;
- whether current booked offense is among the following offenses or is a conspiracy, attempt, solicitation, or FTA of any for those offenses;
  - Murder
  - Voluntary Manslaughter
  - Aggravated Mayhem
  - Torture
  - Felony Sexual Assault
  - Robbery
  - Carjacking
  - Felony Domestic Violence
  - Felony Stalking
  - Violation of a Domestic Violence Protective Order
  - Escape
- whether the current booked offense is deemed violent according to the California PSA List of Violent Offenses and the NVCA flag calculated in step (1) is indicated.

If any of these conditions are true, the individual is automatically given a "Release Not Recommended" (the most restrictive possible recommendation) and the assessment need not continue. This is called a "charge-based exclusion,"

[4] As we describe each component of the version of the PSA as it existed during the period under study, we will highlight analogues to each component under the new version of the PSA.

[5] https://www.psapretrial.org/about/factors.

and we refer to charges that trigger a charge-based exclusion as "exclusion charges." [6] If no charge-based exclusion is made, the assessment continues to step (3).

In step (3), the six point NCA and FTA predictions are combined using the Decision Making Framework (DMF) shown in Figure 1 to arrive at what we refer to as an "initial recommendation."[7] Each cell in the matrix corresponds to a recommended level of supervision. For example, if an individual has an FTA prediction of 2 and a NCA prediction of 3, one would find the entry in the DMF in the second row and third column to arrive at an initial recommendation of OR-NAS, the lowest recommended level of supervision possible.

	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	OR - NAS	OR - NAS				
FTA 2	OR - NAS	OR - NAS	OR - NAS	OR - Minimum	SFPDP - ACM	
FTA 3		OR - NAS	OR - Minimum	SFPDP - ACM	SFPDP - ACM	Release Not Recommended
FTA 4		OR - Minimum	SFPDP - ACM	SFPDP - ACM	Release Not Recommended	Release Not Recommended
FTA 5		SFPDP - ACM	SFPDP - ACM	SFPDP - ACM* / Release Not Recommended	Release Not Recommended	Release Not Recommended
FTA 6				Release Not Recommended	Release Not Recommended	Release Not Recommended

* Release Not Recommended if any booked offense is a felony or violent misdemeanor per PSA Violent Offenses List; SFPDP – ACM if booked offense(s) are non-violent misdemeanors.

Figure 1: Decision Making Framework used in San Francisco pilot study. This matrix translates the empirical risk estimates of NCA and FTA into policy recommendations about the appropriate level of pre-trial supervision for a person with that estimated risk profile.

In this implementation of the tool, the possible levels of supervision in the DMF in order of increasing level of supervision were: (1) No Active Supervision– release on own recognizance and receive court reminders (OR-NAS); (2) Minimum Supervision– release on own recognizance with court reminders and twice weekly phone reporting (OR-Minimum); (3) Assertive Case Management– release on own recognizance or under supervision, including court date reminders, four times weekly reporting with two to four of those times reporting in person, and an out of custody needs assessment (SFPDP-ACM); and, (4) Release not Recommended.[8]

Finally, in step (4), two more determinations are made:

- whether the current booked offense is among the following charges or the current booked offense is a solicitation, conspiracy, attempt or FTA for any of the charges on that list;
  - Violation of other Protective Orders
  - Person to Person Sex Crime
  - Arson
  - Involved the Use of a Weapon, Caustic Chemical, Flammable Substance, or Explosive,
  - Felony inflicting Great Bodily Injury
  - Misdemeanor Domestic Violence
  - Misdemeanor Stalking
- whether the current booked charge is not on the list of violent offenses but the NVCA flag was triggered.

If either of these conditions are true, the initial recommendation is increased one level. For example, if the initial recommendation was OR-NAS, it is increased to OR-Minimum. This recommendation is then the final pre-trial supervision recommendation. [9] We refer to this as a "charge-based bump-up" and charges that trigger a charge-based bump-up as "bump-up charges."

In summary, abandoning jargon and simplifying to the extent possible, we conceptualize the process as follows: an individual's initial recommended level of supervision is determined by combining predictions about their likelihood of FTA and NCA (predictions that do not rely on charge-based information). If there are very serious booking charges (i.e. exclusion charges) or the booking charges are violent and the person is predicted to have a higher likelihood of re-arrest for a violent crime, then the individual is automatically

[6] Under the new version of the PSA, there is no longer a charge-based exclusion as a formal step in the process. The analogue to charge-based exclusions now fall under "additional guidance." For example, in a sample PSA given in [21], the guidebook gives an example of additional guidance to augment the PSA as "If the current charge is a first- or second-degree violent felony, the person may be placed on Release Level 3 (the highest release level), regardless of the PSA scores." This is analogous to a step (2) exclusion under the version of the PSA described here.

[7] This terminology may be confusing, as this comes after step (2) so chronologically in some sense is not initial. By this we mean to say that this is the recommendation that would be given without considering any charge-based amendments to the recommendation. This can be calculated even if an individual has a charge-based exclusion.

[8] What is referred to as the decision-making framework (DMF) in this section has a direct analogue in the new version of the PSA called the Release Conditions Matrix (RCM). One important differentiator between the versions is that in the version described here, the DMF may recommend pre-trial detention. Under the new version with the RCM, according to [21] "Detention is not included in the matrix because eligibility for detention is based on state law, and the matrix becomes relevant only after a judicial officer decides a person will be released."

[9] Similar to step (2), step (4) increases are now included under "additional guidance" rather than as a formalized step. For example, in the sample PSA given in [21], "If there is an NVCA flag, consider increasing the person's release level by one level" is given as an example of additional guidance that could be included. This is analogous to the step (4) increase given here, though less formalized.

recommended for the highest level of supervision. If there are moderately serious booking charges (i.e. bump-up charges) or the individual is predicted to have a higher likelihood of re-arrest for a violent crime despite the current charges not being violent, then the initial recommendation is increased to the next highest level.

4 DATA

In collaboration with the SFPD, we obtained data collected during San Francisco's pilot study of the PSA. This study was conducted over the period from mid-2016 to mid-2017. During this time period, the SFPD saved PSAs for their clients as scanned image files. From the SFPD, we obtained 2450 of these files, the information from which was manually entered into a spreadsheet. The following fields were collected from each PSA: the defendant's unique identification number, name, date of birth, arrest date, date on which the PSA was conducted, NVCA prediction, NCA prediction, FTA prediction, a list of booked charges and corresponding charge codes, the recommendation of the PSA, an indicator of whether the recommendation was the result of charge-based exclusion, and an indicator of whether a charge-based bump-up was applied to the recommendation. We also recorded several of the individual risk factors listed on the PSA for each defendant: age at current offense, prior conviction (misdemeanor or felony), and prior violent conviction.

We separately obtained data from San Francisco's court databases on all defendants who interacted with the court system during the time period of the pilot program. For each defendant, we obtained a unique identification number, name, date of birth, list of booking charges, list of charges filed by the district attorney, and the disposition code for each individual charge. The disposition codes define whether each charge resulted in a conviction and are the basis on which we retrospectively calculate which charges were substantiated and which were not.

5 ANALYSIS

Our analysis consists of three parts: de-duplication and record-linkage; validation; and counterfactual analysis. The goal of the de-duplication and record linkage is to ensure that each arrest only appears once in our dataset and that each arrest record in the PSA data is linked to the correct court case in the court data. This linkage allows us to see whether the booking charges included in each PSA ultimately resulted in a conviction, information that is only contained in the court records. It also provides us additional demographic information about the people who were assessed by the PSA. The result of this step is one unified dataset that contains one copy of each PSA that was administered during the pilot program, the outcome of all charges associated with each PSA, and additional demographic information about the individual to whom each PSA pertains. In completing this process, we drop 64 records due to incompleteness, 419 records as duplicates[10], and 31 records due to an inability to establish a definitive match in the court data. In the end, this leaves us with 1916 records with which to do the analysis. Details of the decision rules used to de-duplicate and match are found in Section 1 of the appendix.

In the validation phase, we create PSA-reproduction code and verify that our code accurately reproduces the outputs of the human-administered PSA when given the same inputs. To do this, we apply our PSA-reproduction code to the linked dataset described above. This code takes the FTA and NCA predictions[11], several variables measuring criminal history, and the booking charges as listed in the court records. It then outputs each of the components of the PSA: a charge-based exclusion indicator, a charge-based bump-up indicator, a violence flag, and a final recommendation. We then compare each of these outputs to those same components as recorded on the original PSA forms. A high rate of agreement assures that our code is accurately representing the PSA.

We find that our calculation of the NVCA flag agrees with that listed on the PSA form for 99.3% of the cases in our dataset. With regard to charge-based exclusions, we find that our reproduction is in agreement with the original PSA data for 99.4% of the records. In the PSA, calculating the charge-based bump-up is not necessary if an exclusion is determined because the recommended level of supervision cannot further be increased. For this reason, we compare our reproduction of charge-based bump-ups to the PSA data only for those records for which a charge-based exclusion was not indicated in the original PSA data. We find that our reproduction is in agreement with the original administration of the PSA for 99.7% of these records. Finally, we compare our calculation of the final recommendation to the recommendation given in the PSA data. We find an agreement of 97.4%, which is slightly lower than for the other components. Nearly all of the disagreements occurred for individuals who fell within the one specific cell of the DMF which required additional determinations to be made (an FTA of 5 and a NCA of 4, shown as the split cell in Figure 1). Based on a manual review, we believe there may have been differing interpretations by the staff administering the PSA as to how these determinations should be applied. In any case, for those cases that do not fall into this cell of the

[10] Many of the dropped records pertain to the same few arrests, which it seems were each saved multiple times.

[11] Because the FTA and NCA predictions do not depend on booking charge, these are held constant and we need not re-calculate them. That is, we do not need to verify that we can reproduce them, as this analysis pertains only to effects driven by perturbations of the booking charges, and these predictions do not change as a function of the booking charges.

DMF, our reproduction of the final recommendation is in agreement with that listed on the PSA form 99.5% of the time. A full discussion of the validation process is given in Section 2 of the appendix.

Finally, having verified that our code accurately reproduces the PSA, we apply our PSA-reproduction code to a counterfactual scenario in which only the charges that resulted in a conviction are used in the calculation of the PSA. This results in two sets of calculations to compare: (a) the PSA's recommendation (and each of its components) based on the booking charges, and (b) the PSA's recommendation (and each of its components) based only on conviction charges. To evaluate the impact that unsubstantiated charges had on the PSA, we compare calculations (a) and (b).

6 RESULTS

In this section, we compare the results of the PSA calculated using the conviction charges to the results of the PSA calculated using the booking charges.[12] Throughout this section, the results presented pertain only to the cases for which all charges had been disposed (or settled) at the time of the analysis, which includes 88.3% of the records.

There is some nuance around which charges should count as convictions. For example, sometimes multiple cases are bundled into a single plea agreement. Should all charges associated with that bundle be counted as conviction charges or only those conviction charges that are part of the case originally associated with the administration of the PSA? Ultimately, we decided that only charges that pertain directly to the arrest that triggered the administration of the PSA ought to be eligible, though we acknowledge that others might disagree with this definition. Thus, for the purposes of this analysis, we define "conviction charges" to be those charges associated with the arrest that triggered the administration of the PSA for which the arrestee was found or plead guilty.[13]

This definition creates some situations where a case outcome indicates that the individual pleaded guilty to other charges, but none of the charges to which they pleaded guilty are associated with the original case. In this scenario, the conviction-charge-PSA is calculated as though there were no charges eligible to trigger charge-based exclusions, bump-ups, or to be considered violent, despite the fact the individual was convicted on some charges (just none that were filed as part of the case associated with their PSA form). We performed additional analysis removing all cases for which a guilty plea was indicated but the individual did not plead guilty to any of the charges associated with the original case. While the exact numbers were lower than those presented in the remainder of this section, qualitatively the results were the same.

Table 1 shows the rate of charge-based exclusions, bump-ups, NVCA flags, and the average recommendation level when calculating each of the components of the PSA. The average recommendation level is based on mapping each level of supervision to its numeric rank: the lowest level of supervision is mapped to one, the second to two, etc. The Charges column gives the charges used to calculate the PSA– either booking charges or conviction charges. The difference between the rate calculated under the booking charges and the rate calculated under the conviction charges is also shown. To test for statistical significance between the components of the PSA under the two input charge conditions, we performed standard statistical hypothesis tests. When comparing the rate of exclusions under booking charges to the rate of exclusions under conviction charges we perform a difference of proportion test. To compare the recommendations, we performed a Wilcoxon rank-sum test, as the recommendations are ordered categorical. Statistical significance at the $\alpha < 0.001$ level is indicated by '$*$' for the differences shown in the results tables.[14] We see that the rate at which charge-based exclusions, bump-ups, and NVCA flags occur is much higher when we consider booking charges relative to when we consider conviction charges as inputs. The average level of recommended pre-trial supervision is also elevated under the calculation using the booking charges relative to that using the conviction charges.

Charges	exclusions	bump-ups	nvca	rec
Conviction	8.7	9.1	9.3	2.5
Booking	29.4	23.7	20.2	3
Difference	20.7 *	14.6 *	10.8 *	0.5 *

Table 1: Percent of cases with exclusions, bump-ups, nvca flags, and the average recommendation by input charges.

Table 2 shows the proportion of people who received a charge-based exclusion, a charge-based bump-up, or an

[12] We do not compare to the original PSA results directly to isolate the effect of altering the input charges. If we compared to the original PSA components, some of the differences we identify may, in fact, be due to some of the differences in interpretation we highlighted in the validation section above.

[13] According to the codebook we received, this is all disposition codes greater than 159. Additionally, by manual review, we have found that cases in which the case is listed as resolved, if some charges associated with a case number have disposition code 72 (plead guilty to other charges) and others charge codes associated with that same case number have disposition code 0, those with disposition code 0 are the ones the individual was convicted of. This was confirmed on several cases by looking at alternative sources of information available in other systems that are not in a database form amenable to statistical analysis.

[14] All p-values are significant at at least the 0.001 level, even after a Bonferroni correction for multiplicity.

NVCA flag when the PSA was calculated using the booking charges but not when it was calculated using the conviction charges. It also shows the percent of people who received a recommendation for more restrictive conditions under the booking charges than under the conviction charges.[15] We find that a substantial portion of the cases (nearly 30%) would have had a lower recommended level of supervision if their PSA had been based only on the charges they were ultimately convicted of.

exclusions	bump-ups	nvca	rec
20.9	17.0	10.9	27.4

Table 2: Percent of cases for which each PSA component was higher under the booking charges than under the conviction charges.

Next we turn to understanding whether overbooking's effect on the PSA differs by race group. For this analysis, we disaggregate the data into two race categories: Black and non-Black. This is an obvious over-simplification, as is any racial categorization. However, based on our analysis of the consistency of racial classification within the court data, we have determined this categorization scheme introduces the fewest problems with inconsistent classification. A full discussion of how we arrived at this decision is available in the appendix.

Table 3 shows equivalent quantities to those shown in Table 1, now disaggregated into the two race groups. We find that overbooking had a larger impact on the rate of charge-based exclusions and the assignment of the NVCA flag for Black people than non-Black people in this data. It had a larger impact on charge-based bump-ups on the non-Black population. However, the impact of overbooking on the ultimate recommendation is, roughly speaking, similar between the two groups. The proportion of cases for which charge-based overrides, bump-ups, NVCA flags were triggered or the recommendation was higher under the conviction charges than under the booking charges is given in Table 4 disaggregated by race. Under this summary of the data, we again see that unsubstantiated charges led to charge-based exclusions at a higher rate for Black defendants than non-Black defendants. However, there is little difference between the groups in terms of the impact of unsubstantiated charges on charge-based bump-ups or on the final recommendation.

To understand how this seemingly paradoxical result is possible, we must first recognize that there are instances

[15]This is different than what is shown in Table 1, as under the former calculation, cases in which the individual was convicted of more severe charges than those under which they were booked offset cases where the reverse occurs.

Charges	group	exclusions	bump-ups	nvca	rec
Conviction	non-black	7.9	8.2	7.7	2.4
Conviction	black	9.7	10.3	11.5	2.6
Booking	non-black	25.9	23.2	15.9	2.9
Booking	black	33.9	24.3	25.7	3.1
Difference	non-black	18 *	15 *	8.2 *	0.5 *
Difference	black	24.2 *	14.1 *	14.2 *	0.5 *

Table 3: Comparison of effects of charges on components of PSA by race for all disposed cases in data

group	exclusions	bump-ups	nvca	rec
non-black	18.1	17.0	8.2	27.5
black	24.5	16.9	14.5	27.2

Table 4: Percent of cases by defendant race with a charge-based exclusion, charge-based bump-up, NVCA flag, or higher recommendations due to unsubstantiated booking charges.

where an individual can have an "unfair" charge-based exclusion or bump-up that does not translate to an "unfair" recommendation. Recall that each of these charge-based components results in an increase to the recommended level of supervision above and beyond the initial recommendation.

Consider, for example, an individual whose initial recommendation is the highest level and who has exclusion or bump-up charges at booking that they are not convicted of. This individual would be classified as having an unfair exclusion or bump-up. However, because their initial recommendation was maximal, whether we calculate the PSA using the booking charges (which would include an exclusion or a bump-up) or we calculate it using only the conviction charges (which would not include an exclusion or bump-up), the recommendation is the same. In the former case, the initial recommendation was maximal and applying the exclusion or bump-up did not increase the recommendation, as it could not further increase. In the latter case, we do not apply the exclusion or bump-up, and the recommendation is still the highest category. Thus, even if exclusion or bump-up booking charges are unsubstantiated, the ultimate recommendation does not change based on those charges for people whose initial recommendation is the highest level. Individuals in this category contribute to the disparity shown under exclusions and bump-ups in Tables 3 and 4 and do not contribute to any difference in recommendations.

Similarly, consider a second scenario where an individual has an initial recommendation of SFPDP-ACM, the second highest level of supervision. If this person is booked under an exclusion charge that is reduced to a bump-up charge that they are convicted of, in both cases, the final recommendation will be the highest level of supervision. To break

this down further, under the booking charges, they receive an exclusion and are automatically moved to the highest category, Release Not Recommended. Under the conviction charges, they receive a charge-based bump-up, which because they began in the second-highest category, also results in a Release Not Recommended recommendation. Thus, under both the booking charges and the conviction charges, their recommendation will be the same, though they will still be classified as having had an unfair exclusion.

Both scenarios where unfair exclusion charges do not materialize into unfair recommendations are only possible when the individual has an initial recommendation that is either the highest category or the second highest category. In the population examined here, the distribution of initial scores was shifted higher for the Black individuals than the non-Black individuals. See Figure 2, which shows the distribution of initial scores broken down into Black and non-Black people. Depending on the definition of fairness adopted, this group-wise distributional difference may itself be indicative of unfairness in the model. However, because our goal is to study the effect of overbooking in isolation, we do not further delve into this other than to note this disparity in the rate at which Black versus non-back people are recommended for pre-trial detention. Regardless, because Black defendants were more likely to fall into the highest or second highest category before any charge-based amendments were made, Black defendants who had unfair charge-based exclusions were more likely to not have those unfair charge-based exclusions impact their final recommendation.

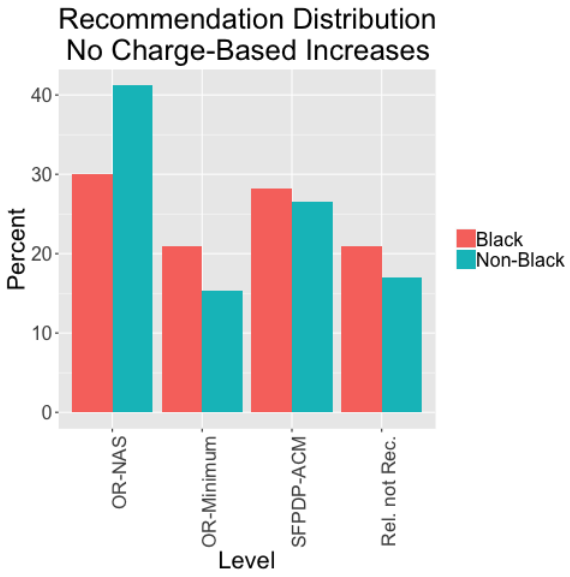

Figure 2: Distribution of initial recommendation, disaggregated into Black and non-Black people.

It is important to note that this conclusion only holds for this particular DMF, the matrix that translates raw predictions of FTA and NCA into initial recommendations. In a jurisdiction with a different DMF under which fewer combinations of risk scores correspond to the highest or second-highest level of supervision, it is possible that the racial disparities in unwarranted charge-based exclusions, bump-ups, and NVCA flags might, in fact, translate to disparities in the recommendations as well. Thus, jurisdictions seeking to amend their DMF (or similarly, their release recommendation matrices) should be aware of this possibility: changes to the release recommendations that are intended to result in a greater rate of release may introduce a racial skew in the impact of overbooking where one previously did not exist.

7 LIMITATIONS

Perhaps most obviously, one limitation of this analysis is that it was done using a fairly small sample size with a limited scope of the cases it covers. This sample may not include all of the PSAs administered during the time period. In order to generalize the findings here, one would need to repeat the analysis using data from a broader geographic range and preferably across a longer time horizon.

Additionally, as with many endeavors to measure criminality, our measure is imperfect. As discussed in the introduction, higher booking charges may translate to higher conviction charges, as later stages of the criminal justice process anchoring to the booking charges. This would cause our analysis to understate the impact of overbooking. The other side of the argument is that because we have excluded conviction charges that are associated with other cases, even when the individual pleaded guilty to those charges as part of a joint deal covering both the case we consider and the case containing the conviction charges, we are understating the extent of guilty charges that ought to count. This would lead to us overstating the extent of overbooking. In the end, there is no perfect measure of criminal behavior, and we believe the measure we have chosen is reasonable.

Finally, we have analyzed the risk assessment in isolation, rather than taking a more holistic approach to analyzing its role within the larger system. For example, we have presented no analysis of the impact of overbooking on judicial decision-making nor analyzed the effect on downstream outcomes like recidivism. In order to understand the real world consequences of overbooking on the individuals evaluated by the risk assessment, more investigation is needed. In this vein, the observed lack of racial disparity in the extent to which overbooking impacts the tool's final recommendation should not be taken as definitive proof that overbooking does not have a racially disparate impact on judicial decision-making. Each component of the PSA we have evaluated is

displayed on the sheet available to judges. It is possible that the presence of a charge-based exclusion, bump-up, or NVCA flag influences the judge's decision-making independent of the tool's final recommendation. If this is so, then the finding that overbooking impacted charge-based exclusions at a higher rate for Black than non-Black individuals, for example, might actually result in a meaningful difference in terms of the impact of overbooking on the judge's decision. This despite the fact that we observed no difference between race groups in the impact of overbooking on the final recommendation of the tool. Without data on what decisions judges made when presented with these evaluations, we cannot say what impact overbooking had on the decisions at an aggregate level or disaggregated by race.

8 CONCLUSION

We have analyzed the effect of overbooking on a popular risk assessment tool, Arnold Ventures' PSA. We have found that for around 27% of the cases analyzed, charges for which the person was not ultimately convicted caused the tool to issue a recommendation for pre-trial supervision that was more restrictive than would have been issued had such charges not been included as inputs to the model. At a more granular level, we have found that a significant portion of the population that was evaluated by the tool, between 10-20%, received charge-based exclusions, charge-based bump-ups, and NVCA flags also based upon charges that were ultimately unsubstantiated by the courts. Disaggregating the analysis by race shows that while Black individuals received unwarranted charge-based exclusions and NVCA flags at a higher rate than non-Black individuals, they did not receive increased recommendations at a substantially higher rate due to the fact that Black individuals were more likely to be classified in the higher risk groups even before charge-based increases are applied. This finding in and of itself may be worrisome to those who hope that risk assessment will close the racial gap in criminal justice outcomes.

This work also naturally raises the question of how one might protect against charges that will ultimately not result in convictions impacting an individual's pre-trial release recommendation. It is not possible to wait until a case is resolved to use only conviction charges to compute a recommendation for pre-trial release conditions, as by definition, by that time the pre-trial period has come to a close. One possible mitigation is to have a defendant advocate present earlier in the process to help catch charges for which the early evidence is lacking. This may be difficult in practice, as the police report is typically not written at the time the PSA is administered. Without the police report, a defendant advocate may not have sufficient information to assess the strength of the evidence for each of the charges.

This work also reveals one possibility for gaming the risk assessment. In theory, in cases where there is ambiguity as to which charges are appropriate for a given observed behavior, a booking officer could opt to book under more serious charges with the intention of inducing a more restrictive detention decision. This could be monitored by tracking the ratio of similar higher and lower charges over time and across protected group status. A heightened ratio may indicate strategic behavior intended to prompt higher recommendations by the tool.

This possibility points to the fragility of such models under highly discretionary inputs. In our view, this raises questions about the appropriateness of their use in high stakes settings without adequate controls and accountability for the inputs. In particular, we believe this reveals the importance of increased accountability and feedback for officers who systematically overbook. This is especially important if booking charges are to be legitimized by their use as inputs to the tool, the output of which— though directly dependent on booking charges— may not be viewed with the same level of skepticism as the charges themselves might be. One potentially revealing line of future research would be to disaggregate the impact of overbooking by booking officer as a means to determine if particular officers' booking decisions are routinely unfairly impacting the tool's pre-trial supervision recommendations. Holding such individuals accountable for their booking decisions may encourage more conservative decisions around booking arrested people on heightened charges, reducing the impact of overbooking even beyond the risk assessment.

Finally, while judges have the power to set more restrictive conditions of release for people booked on certain serious charges, these types of charges do not necessarily correlate with an increased risk of undesirable outcomes. Embedding charge-based considerations within the final recommendation may communicate to judges that merely having been booked on such charges increases an individual's empirical risk level when, in fact, such overrides are an instantiation of a policy driven by politics rather than being directly tied to an individual's risk. Given the large impact of charging decisions on the output of the tool, it is critical that judges understand this distinction in order to properly discount increased recommendations that are based on questionably or weakly supported charges.

ACKNOWLEDGMENTS

We are grateful to Chelsea Barabas, Logan Koepke, Alicia Solow-Niederman, and David Robinson for helpful comments on an early draft. This work was supported by The Ethics and Governance of AI Initiative The MacArthur Foundation, and the Ford Foundation.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).

[2] 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).

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

[4] 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. (2018).

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

[6] Will Dobbie, Jacob Goldin, and Crystal S Yang. 2018. The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review* 108, 2 (2018), 201–40.

[7] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.

[8] Arpit Gupta, Christopher Hansman, and Ethan Frenchman. 2016. The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies* 45, 2 (2016), 471–505.

[9] Bernard E Harcourt. 2015. Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter* 27, 4 (2015), 237–243.

[10] 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.

[11] Jon Kleinberg. 2018. Inherent Trade-Offs in Algorithmic Fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '18)*. ACM, New York, NY, USA, 40–40.

[12] Megan C Kurlychek and Brian D Johnson. 2019. Cumulative Disadvantage in the American Criminal Justice System. *Annual Review of Criminology* 2 (2019), 291–319.

[13] Emily Leslie and Nolan G Pope. 2017. The unintended impact of pretrial detention on case outcomes: Evidence from New York City arraignments. *The Journal of Law and Economics* 60, 3 (2017), 529–557.

[14] Kristian Lum, Erwin Ma, and Mike Baiocchi. 2017. The causal impact of bail on case outcomes for indigent defendants in New York City. *Observational Studies* 3 (2017), 39–64.

[15] Sandra G. Mayson. 2019. Bias In, Bias Out. *The Yale Law Journal* 128, 8 (2019), 2122–2473.

[16] Emily Owens, Erin M Kerrison, and B Santos Da Silveira. 2017. Examining racial disparities in criminal case outcomes among indigent defendants in San Francisco. *Quattrone Center for the Fair Administration of Justice, University of Pennsylvania Law School* (2017).

[17] Cindy Redcross, Brit Henderson, Luke Miratrix, and Erin Valentine. 2019. *Evaluation of Pretrial Justice System Reforms that Use the Public Safety Assessment*. Technical Report. MDRC Center for Criminal Justice Research.

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

[19] Alicia Solow-Niederman, YooJung Choi, and Guy Van den Broeck. [n. d.]. The Institutional Life of Algorithmic Risk Assessment. *Berkeley Technology Law Journal* ([n. d.]).

[20] Megan T Stevenson. 2018. Distortion of justice: How the inability to pay bail affects case outcomes. *The Journal of Law, Economics, and Organization* 34, 4 (2018), 511–542.

[21] Arnold Ventures. [n. d.]. *Guide to the Release Conditions Matrix*.