CORRECTED 30 MARCH 2018; SEE FULL TEXT

SCIENCE ADVANCES | RESEARCH ARTICLE

RESEARCH METHODS

The accuracy, fairness, and limits of predicting recidivism

Julia Dressel and Hany Farid*

Algorithms for predicting recidivism are commonly used to assess a criminal defendant's likelihood of committing a crime. These predictions are used in pretrial, parole, and sentencing decisions. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. We show, however, that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear predictor with only two features.

INTRODUCTION

We are the frequent subjects of predictive algorithms that determine music recommendations, product advertising, university admission, job placement, and bank loan qualification. In the criminal justice system, predictive algorithms have been used to predict where crimes will most likely occur, who is most likely to commit a violent crime, who is likely to fail to appear at their court hearing, and who is likely to reoffend at some point in the future (*1*).

One widely used criminal risk assessment tool, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS; Northpointe, which rebranded itself to "equivant" in January 2017), has been used to assess more than 1 million offenders since it was developed in 1998. The recidivism prediction component of COMPAS—the recidivism risk scale—has been in use since 2000. This software predicts a defendant's risk of committing a misdemeanor or felony within 2 years of assessment from 137 features about an individual and the individual's past criminal record.

Although the data used by COMPAS do not include an individual's race, other aspects of the data may be correlated to race that can lead to racial disparities in the predictions. In May 2016, writing for *ProPublica*, Angwin *et al.* (*2*) analyzed the efficacy of COMPAS on more than 7000 individuals arrested in Broward County, Florida between 2013 and 2014. This analysis indicated that the predictions were unreliable and racially biased. COMPAS's overall accuracy for white defendants is 67.0%, only slightly higher than its accuracy of 63.8% for black defendants. The mistakes made by COMPAS, however, affected black and white defendants differently: Black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9%, nearly twice as high as their white counterparts at 23.5%; and white defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7%, nearly twice as high as their black counterparts at 28.0%. In other words, COMPAS scores appeared to favor white defendants over black defendants by underpredicting recidivism for white and overpredicting recidivism for black defendants.

In response to this analysis, Northpointe argued that the *ProPublica* analysis overlooked other more standard measures of fairness that the COMPAS score satisfies (*3*) [see also the studies of Flores *et al.* (*4*) and Kleinberg *et al.* (*5*)]. Specifically, it is argued that the COMPAS score is not biased against blacks because the likelihood of recidivism among high-risk offenders is the same regardless of race (predictive parity), it can discriminate between recidivists and nonrecidivists equally well for white and black defendants as measured with the area under the curve of the receiver operating characteristic, AUC-ROC (accuracy equity), and the likelihood of recidivism for any given score is the same regardless of race (calibration). The disagreement amounts to different definitions of fairness. In an eloquent editorial, Corbett-Davies *et al.* (*6*) explain that it is impossible to simultaneously satisfy all of these definitions of fairness because black defendants have a higher overall recidivism rate (in the Broward County data set, black defendants recidivate at a rate of 51% as compared with 39% for white defendants, similar to the national averages).

While the debate over algorithmic fairness continues, we consider the more fundamental question of whether these algorithms are any better than untrained humans at predicting recidivism in a fair and accurate way. We describe the results of a study that shows that people from a popular online crowdsourcing marketplace—who, it can reasonably be assumed, have little to no expertise in criminal justice—are as accurate and fair as COMPAS at predicting recidivism. In addition, although Northpointe has not revealed the inner workings of their recidivism prediction algorithm, the accuracy of COMPAS on one data set can be explained with a simple classifier (*7*); we confirm their result here. In further agreement with Angelino *et al.* (*7*), we also show that although COMPAS may use up to 137 features to make a prediction, the same predictive accuracy can be achieved with only two features, and that more sophisticated classifiers do not improve prediction accuracy or fairness. Collectively, these results cast significant doubt on the entire effort of algorithmic recidivism prediction.

RESULTS

We compare the overall accuracy and bias in human assessment with the algorithmic assessment of COMPAS. Throughout, a positive prediction is one in which a defendant is predicted to recidivate, whereas a negative prediction is one in which they are predicted to not recidivate. We measure overall accuracy as the rate at which a defendant is correctly predicted to recidivate or not (that is, the combined true-positive and true-negative rates). We also report on false positives (a defendant is predicted to recidivate but they do not) and false negatives (a defendant is predicted to not recidivate but they do).

Human assessment

Participants saw a short description of a defendant that included the defendant's sex, age, and previous criminal history, but not their race (see Materials and Methods). Participants predicted whether this person would recidivate within 2 years of their most recent crime. We used

6211 Sudikoff Laboratory, Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA.
*Corresponding author. Email: farid@dartmouth.edu

a total of 1000 defendant descriptions that were randomly divided into 20 subsets of 50 each. To make the task manageable, each participant was randomly assigned to see one of these 20 subsets. The mean and median accuracy for these predictions is 62.1 and 64.0%.

We compare these results with the performance of COMPAS on this subset of 1000 defendants. Because groups of 20 participants judged the same subset of 50 defendants, the individual judgments are not independent. However, because each participant judged only one subset of the defendants, the median accuracies of each subset can reasonably be assumed to be independent. Therefore, the participant performance on the 20 subsets can be directly compared to the COMPAS performance on the same 20 subsets. A one-sided *t* test reveals that the average of the 20 median participant accuracies of 62.8% [and a standard deviation (SD) of 4.8%] is, just barely, lower than the COMPAS accuracy of 65.2% ($P = 0.045$).

To determine whether there is "wisdom in the crowd" (*8*) (in our case, a small crowd of 20 per subset), participant responses were pooled within each subset using a majority rules criterion. This crowd-based approach yields a prediction accuracy of 67.0%. A one-sided *t* test reveals that COMPAS is not significantly better than the crowd ($P = 0.85$).

Prediction accuracy can also be assessed using the AUC-ROC. The AUC-ROC for our participants is 0.71 ± 0.03, nearly identical to COMPAS's 0.70 ± 0.04.

Prediction accuracy can also be assessed using tools from signal detection theory in which accuracy is expressed in terms of sensitivity (d') and bias (β). Higher values of d' correspond to greater participant sensitivity. A value of $d' = 0$ means that the participant has no information to make reliable identifications no matter what bias he or she might have. A value of $\beta = 1.0$ indicates no bias, a value of $\beta > 1$ indicates that participants are biased to classifying a defendant as not being at risk of recidivating, and $\beta < 1$ indicates that participants are biased to classifying a defendant as being at risk of recidivating. With a d' of 0.86 and a β of 1.02, our participants are slightly more sensitive and slightly less biased than COMPAS with a d' of 0.77 and a β of 1.08.

With considerably less information than COMPAS (only 7 features compared to COMPAS's 137), a small crowd of nonexperts is as accurate as COMPAS at predicting recidivism. In addition, our participants' and COMPAS's predictions were in agreement for 692 of the 1000 defendants.

Fairness

We measure the fairness of our participants with respect to a defendant's race based on the crowd predictions. Our participants' accuracy on black defendants is 68.2% compared with 67.6% for white defendants. An unpaired *t* test reveals no significant difference across race ($P = 0.87$). This is similar to that of COMPAS that has an accuracy of 64.9% for black defendants and 65.7% for white defendants, which is also not significantly different ($P = 0.80$, unpaired *t* test). By this measure of fairness, our participants and COMPAS are fair to black and white defendants.

Our participants' false-positive rate for black defendants is 37.1% compared with 27.2% for white defendants. An unpaired *t* test reveals a significant difference across race ($P = 0.027$). Our participants' false-negative rate for black defendants is 29.2% compared with 40.3% for white defendants. An unpaired *t* test reveals a significant difference across race ($P = 0.034$). These discrepancies are similar to that of COMPAS that has a false-positive rate of 40.4% for black defendants and 25.4% for white defendants, which are significantly different ($P = 0.002$, unpaired *t* test). COMPAS's false-negative rate for black defendants is 30.9% compared with 47.9% for white defendants, which are significantly different ($P = 0.003$, unpaired *t* test). By this measure of fairness, our participants and COMPAS are similarly unfair to black defendants, despite the fact that race is not explicitly specified. See Table 1 [columns (A) and (C)] and Fig. 1 for a summary of these results.

Prediction with race

In this second condition, a newly recruited set of 400 participants repeated the same study but with the defendant's race included. We wondered whether including a defendant's race would reduce or exaggerate the effect of any implicit, explicit, or institutional racial bias. In this condition, the mean and median accuracy on predicting whether a defendant would recidivate is 62.3 and 64.0%, nearly identical to the condition where race is not specified.

The crowd-based accuracy is 66.5%, slightly lower than the condition where race is not specified, but not significantly different ($P = 0.66$, paired *t* test). The crowd-based AUC-ROC is 0.71 ± 0.03 and the d'/β is 0.83/1.03, similar to the previous no-race condition [Table 1, columns (A) and (B)].

With respect to fairness, participant accuracy is not significantly different for black defendants (66.2%) compared with white defendants (67.6%; $P = 0.65$, unpaired *t* test). The false-positive rate for black defendants is 40.0% compared with 26.2% for white defendants. An unpaired *t* test reveals a significant difference across race ($P = 0.001$). The false-negative rate for black defendants is 30.1% compared with 42.1% for white defendants that, again, is significantly different ($P = 0.030$, unpaired *t* test). See Table 1 [column (B)] for a summary of these results.

In conclusion, there is no sufficient evidence to suggest that including race has a significant impact on overall accuracy or fairness. The exclusion of race does not necessarily lead to the elimination of racial disparities in human recidivism prediction.

Participant demographics

Our participants ranged in age from 18 to 74 (with one participant over the age of 75) and in education level from "less than high school degree" to "professional degree." Neither age, gender, nor level of education had a significant effect on participant accuracy. There were not enough nonwhite participants to reliably measure any differences across participant race.

Table 1. Human versus COMPAS algorithmic predictions from 1000 defendants. Overall accuracy is specified as percent correct, AUC-ROC, and criterion sensitivity (d') and bias (β). See also Fig. 1.

	(A) Human (no race)	(B) Human (race)	(C) COMPAS
Accuracy (overall)	67.0%	66.5%	65.2%
AUC-ROC (overall)	0.71	0.71	0.70
d'/β (overall)	0.86/1.02	0.83/1.03	0.77/1.08
Accuracy (black)	68.2%	66.2%	64.9%
Accuracy (white)	67.6%	67.6%	65.7%
False positive (black)	37.1%	40.0%	40.4%
False positive (white)	27.2%	26.2%	25.4%
False negative (black)	29.2%	30.1%	30.9%
False negative (white)	40.3%	42.1%	47.9%

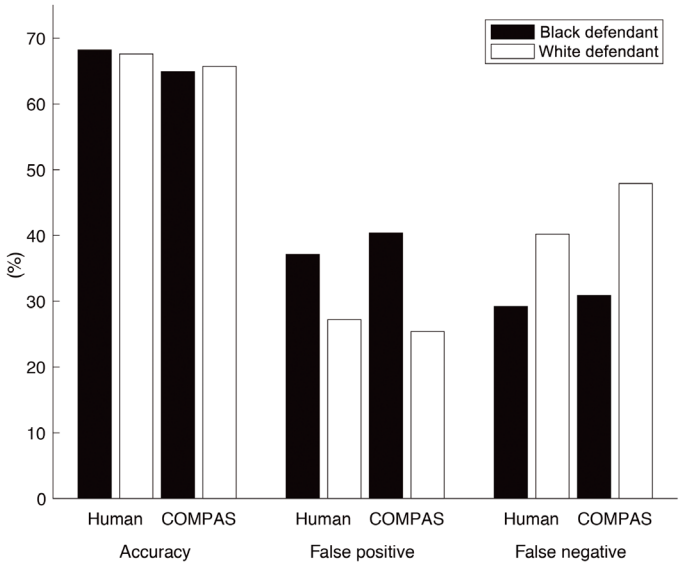

Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

Algorithmic assessment

Because nonexperts are as accurate as the COMPAS software, we wondered about the sophistication of the COMPAS predictive algorithm. Northpointe's COMPAS software incorporates 137 distinct features to predict recidivism. With an overall accuracy of around 65%, these predictions are not as accurate as we might want, particularly from the point of view of a defendant whose future lies in the balance.

Northpointe does not reveal the details of the inner workings of COMPAS—understandably so, given their commercial interests. We have, however, found that a simple linear predictor—logistic regression (LR) (see Materials and Methods)—provided with the same seven features as our participants (in the no-race condition), yields similar prediction accuracy as COMPAS. As compared to COMPAS's overall accuracy of 65.4%, the LR classifier yields an overall testing accuracy of 66.6%. This predictor yields similar results to COMPAS in terms of predictive fairness [Table 2, (A) and (D) columns].

Despite using only 7 features as input, a standard linear predictor yields similar results to COMPAS's predictor with 137 features. We can reasonably conclude that COMPAS is using nothing more sophisticated than a linear predictor or its equivalent.

To test whether performance was limited by the classifier or by the nature of the data, we trained a more powerful nonlinear support vector machine (NL-SVM) on the same data. Somewhat surprisingly, the NL-SVM yields nearly identical results to the linear classifier [Table 2, column (C)]. If the relatively low accuracy of the linear classifier was because the data are not linearly separable, then we would have expected the NL-SVM to perform better. The failure to do so suggests that the data are simply not separable, linearly, or otherwise.

Lastly, we wondered whether using an even smaller subset of the 7 features would be as accurate as using COMPAS's 137 features. We trained and tested an LR classifier on all possible subsets of the seven features. A classifier based on only two features—age and total number of previous convictions—performs as well as COMPAS; see Table 2 [column (B)]. The importance of these two criteria is consistent with the conclusions of two meta-analysis studies that set out to determine, in part, which criteria are most predictive of recidivism (*9*, *10*).

DISCUSSION

We have shown that commercial software that is widely used to predict recidivism is no more accurate or fair than the predictions of people with little to no criminal justice expertise who responded to an online survey. Given that our participants, our classifiers, and COMPAS all seemed to reach a performance ceiling of around 65% accuracy, it is important to consider whether any improvement is possible. We should note that our participants were each presented with the same data for each defendant and were not instructed on how to use these data in making a prediction. It remains to be seen whether their prediction accuracy would improve with the addition of guidelines that specify how much weight individual features should be given. For example, a large-scale meta-analysis of approaches to predicting recidivism of sexual offenders (*11*) found that actuarial measures, in which explicit data and explicit combination rules are used to combine the data into a single score, provide more accurate predictions than unstructured measures in which neither explicit data nor explicit combination rules are specified. It also remains to be seen whether the addition of dynamic risk factors (for example, pro-offending attitudes and socio-affective problems) would improve prediction accuracy as previously suggested (*12*, *13*) (we note, however, that COMPAS does use some dynamic risk factors that do not appear to improve overall accuracy). Lastly, because pooling responses from multiple participants yields higher accuracy than individual responses, it remains to be seen whether a larger pool of participants will yield even higher accuracy, or whether participants with criminal justice expertise would outperform those without.

Although Northpointe does not reveal the details of their COMPAS software, we have shown that their prediction algorithm is equivalent to a simple linear classifier. In addition, despite the impressive collection of 137 features, it would appear that a linear classifier based on only 2 features—age and total number of previous convictions—is all that is required to yield the same prediction accuracy as COMPAS. This finding is supported by earlier work that showed how to create simple, transparent, and interpretable predictive rules for recidivism prediction (*7*, *14*, *15*). When applied to the same Broward County data set, the authors found that simple models afforded similar predictive accuracy as COMPAS.

The question of accurate prediction of recidivism is not limited to COMPAS. A review of nine different algorithmic approaches to predicting recidivism found that eight of the nine approaches failed to make accurate predictions (including COMPAS) (*16*). In addition, a meta-analysis of nine algorithmic approaches found only moderate levels of predictive accuracy across all approaches and concluded that these techniques should not be solely used for criminal justice decision-making, particularly in decisions of preventative detention (*17*).

Recidivism in this study, and for the purpose of evaluating COMPAS, is operationalized with rearrest that, of course, is not a direct measure of reoffending. As a result, differences in the arrest rate of black and white defendants complicate the direct comparison of false-positive and false-negative rates across race (black people, for example, are almost four times as likely as white people to be arrested for drug offenses).

When considering using software such as COMPAS in making decisions that will significantly affect the lives and well-being of criminal defendants, it is valuable to ask whether we would put these decisions in the hands of random people who respond to an online survey because, in the end, the results from these two approaches appear to be indistinguishable.

Table 2. Algorithmic predictions from 7214 defendants. Logistic regression with 7 features (A) (LR_7), logistic regression with 2 features (B) (LR_2), a nonlinear SVM with 7 features (C) (NL-SVM), and the commercial COMPAS software with 137 features (D) (COMPAS). The results in columns (A), (B), and (C) correspond to the average testing accuracy over 1000 random 80%/20% training/testing splits. The values in the square brackets correspond to the 95% bootstrapped [columns (A), (B), and (C)] and binomial [column (D)] confidence intervals.

	(A) LR_7	(B) LR_2	(C) NL-SVM	(D) COMPAS
Accuracy (overall)	66.6% [64.4, 68.9]	66.8% [64.3, 69.2]	65.2% [63.0, 67.2]	65.4% [64.3, 66.5]
Accuracy (black)	66.7% [63.6, 69.6]	66.7% [63.5, 69.2]	64.3% [61.1, 67.7]	63.8% [62.2, 65.4]
Accuracy (white)	66.0% [62.6, 69.6]	66.4% [62.6, 70.1]	65.3% [61.4, 69.0]	67.0% [65.1, 68.9]
False positive (black)	42.9% [37.7, 48.0]	45.6% [39.9, 51.1]	31.6% [26.4, 36.7]	44.8% [42.7, 46.9]
False positive (white)	25.3% [20.1, 30.2]	25.3% [20.6, 30.5]	20.5% [16.1, 25.0]	23.5% [20.7, 26.5]
False negative (black)	24.2% [20.1, 28.2]	21.6% [17.5, 25.9]	39.6% [34.2, 45.0]	28.0% [25.7, 30.3]
False negative (white)	47.3% [40.8, 54.0]	46.1% [40.0, 52.7]	56.6% [50.3, 63.5]	47.7% [45.2, 50.2]

MATERIALS AND METHODS

Our analysis was based on a database of 2013–2014 pretrial defendants from Broward County, Florida (*2*). This database of 7214 defendants contains individual demographic information, criminal history, the COMPAS recidivism risk score, and each defendant's arrest record within a 2-year period following the COMPAS scoring. COMPAS scores, ranging from 1 to 10, classify the risk of recidivism as low-risk (1 to 4), medium-risk (5 to 7), or high-risk (8 to 10).

Our algorithmic assessment was based on this full set of 7214 defendants. Our human assessment was based on a random subset of 1000 defendants, which was held fixed throughout all conditions. This subset yielded similar overall COMPAS accuracy, false-positive rate, and false-negative rate as the complete database (a positive prediction is one in which a defendant is predicted to recidivate; a negative prediction is one in which they are predicted to not recidivate). The COMPAS accuracy for this subset of 1000 defendants was 65.2%. The average COMPAS accuracy on 10,000 random subsets of size 1000 each was 65.4% with a 95% confidence interval of (62.6, 68.1).

Human assessment

A descriptive paragraph for each of 1000 defendants was generated:

> The defendant is a [SEX] aged [AGE]. They have been charged with: [CRIME CHARGE]. This crime is classified as a [CRIMINAL DEGREE]. They have been convicted of [NON-JUVENILE PRIOR COUNT] prior crimes. They have [JUVENILE- FELONY COUNT] juvenile felony charges and [JUVENILE-MISDEMEANOR COUNT] juvenile misdemeanor charges on their record.

In a follow-up condition, the defendant's race was included so that the first line of the above paragraph read, "The defendant is a [RACE] [SEX] aged [AGE]."

There were a total of 63 unique criminal charges including armed robbery, burglary, grand theft, prostitution, robbery, and sexual assault. The crime degree is either "misdemeanor" or "felony." To ensure that our participants understood the nature of each crime, the above paragraph was followed by a short description of each criminal charge:

[CRIME CHARGE]: [CRIME DESCRIPTION]

After reading the defendant description, participants were then asked to respond either "yes" or "no" to the question "Do you think this person will commit another crime within 2 years?" The participants were required to answer each question and could not change their response once it was made. After each answer, the participants were given two forms of feedback: whether their response was correct and their average accuracy.

The 1000 defendants were randomly divided into 20 subsets of 50 each. Each participant was randomly assigned to see one of these 20 subsets. The participants saw the 50 defendants, one at a time, in random order. The participants were only allowed to complete a single subset of 50 defendants.

The participants were recruited through Amazon's Mechanical Turk, an online crowdsourcing marketplace where people are paid to perform a wide variety of tasks (Institutional Review Board guidelines were followed for all participants). Our task was titled "Predicting Crime" with the description "Read a few sentences about an actual person and predict if they will commit a crime in the future." The keywords for the task were "survey, research, and criminal justice." The participants were paid $1.00 for completing the task and a $5.00 bonus if their overall accuracy on the task was greater than 65%. This bonus was intended to provide an incentive for participants to pay close attention to the task. To filter out participants who were not paying close attention, three catch trials were randomly added to the subset of 50 questions. These questions were formatted to look like all other questions but had easily identifiable correct answers. A participant's response was eliminated from our analysis if any of these questions were answered incorrectly. The catch trial questions were (i) The state of California was the 31st state to join the Union. California's nickname is: The Golden State. The state capital is Sacramento. California is bordered by three other states. Los Angeles is California's most populous city, which is the country's second largest city after New York City. Does the state of California have a nickname?; (ii) The first spaceflight that landed humans on the Moon was Apollo 11. These humans were: Neil Armstrong and Buzz Aldrin. Armstrong was the first person to step onto the lunar surface. This landing occurred in 1969. They collected 47.5 pounds (21.59 kg) of lunar material to bring back to Earth. Did the first spaceflight that landed humans on the Moon carry Buzz Aldrin?; and (iii) The Earth is the third planet from the Sun. The shape of Earth is approximately

oblate spheroidal. It is the densest planet in the Solar System and the largest of the four terrestrial planets. During one orbit around the Sun, Earth rotates about its axis over 365 times. Earth is home to over 7.4 billion humans. Is Earth the fifth planet from the Sun?

Responses for the first (no-race) condition were collected from 462 participants, 62 of which were removed because of an incorrect response on a catch trial. Responses for the second (race) condition were collected from 449 participants, 49 of which were removed because of an incorrect response on a catch trial. In each condition, this yielded 20 participant responses for each of 20 subsets of 50 questions. Because of the random pairing of participants to a subset of 50 questions, we occasionally oversampled the required number of 20 participants. In these cases, we selected a random 20 participants and discarded any excess responses. Throughout, we used both paired and unpaired *t* tests (with 19 degrees of freedom) to analyze the performance of our participants and COMPAS.

Algorithmic assessment

Our algorithmic analysis used the same seven features as described in the previous section extracted from the records in the Broward County database. Unlike the human assessment that analyzed a subset of these defendants, the following algorithmic assessment was performed over the entire database.

We used two different classifiers: logistic regression (*18*) (a linear classifier) and a nonlinear SVM (*19*). The input to each classifier was seven features from 7214 defendants: age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior (nonjuvenile) crimes, crime degree, and crime charge (see previous section). Each classifier was trained to predict recidivism from these seven features. Each classifier was trained 1000 times on a random 80% training and 20% testing split; we report the average testing accuracy and bootstrapped 95% confidence intervals for these classifiers.

Logistic regression is a linear classifier that, in a two-class classification (as in our case), computes a separating hyperplane to distinguish between recidivists and nonrecidivists. A nonlinear SVM uses a kernel function—in our case, a radial basis kernel—to project the initial seven-dimensional feature space to a higher dimensional space in which a linear hyperplane is used to distinguish between recidivists and nonrecidivists. The use of a kernel function amounts to computing a nonlinear separating surface in the original seven-dimensional feature space, allowing the classifier to capture more complex patterns between recidivists and nonrecidivists than is possible with linear classifiers.

REFERENCES AND NOTES

1. W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, J. S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (Rand Corporation, 2013).
2. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, 23 May 2016; www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
3. W. Dieterich, C. Mendoza, T. Brennan, "COMPAS risk scales: Demonstrating accuracy equity and predictive parity" (Technical Report, Northpointe Inc., 2016).
4. A. W. Flores, K. Bechtel, C. T. Lowenkamp, False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *Fed. Prob.* **80**, 38 (2016).
5. J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores; https://arxiv.org/abs/1609.05807v2 (2016).
6. S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear," *Washington Post*, 17 October 2016; www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas.
7. E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning Certifiably Optimal Rule Lists, in *2017 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2017), pp. 35–44.
8. R. Hastie, T. Kameda, The robust beauty of majority rules in group decisions. *Psychol. Rev.* **112**, 494–508 (2005).
9. P. Gendreau, T. Little, C. Goggin, A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology* **34**, 575–608 (1996).
10. R. K. Hanson, M. T. Bussière, Predicting relapse: A meta-analysis of sexual offender recidivism studies. *J. Consult. Clin. Psychol.* **66**, 348–362 (1998).
11. R. K. Hanson, K. E. Morton-Bourgon, The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychol. Assess.* **21**, 1–21 (2009).
12. R. K. Hanson, A. J. Harris, Where should we intervene? Dynamic predictors of sexual offense recidivism. *Crim. Justice Behav.* **27**, 6–35 (2000).
13. A. Beech, C. Friendship, M. Erikson, R. K. Hanson, The relationship between static and dynamic risk factors and reconviction in a sample of U.K. child abusers. *Sex. Abuse* **14**, 155–167 (2002).
14. J. Zeng, B. Ustun, C. Rudin, Interpretable classification models for recidivism prediction. *J. R. Stat. Soc. A* **180**, 689–722 (2017).
15. E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, C. Rudin, Learning Certifiably Optimal Rule Lists for Categorical Data, arXiv:1704.01701 (2017).
16. K. A. Geraghty, J. Woodhams, The predictive validity of risk assessment tools for female offenders: A systematic review. *Aggress. Violent Behav.* **21**, 25 (2015).
17. M. Yang, S. C. Wong, J. Coid, The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychol. Bull.* **136**, 740–767 (2010).
18. D. R. Cox, The regression analysis of binary sequences. *J. R. Stat. Soc. B* **20**, 215–242 (1958).
19. C. Cortes, V. Vapnik, Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

Acknowledgments: We wish to thank M. Banks, M. Bravo, E. Cooper, L. Lax, and G. Wolford for helpful discussions. **Funding:** The authors acknowledge that they received no funding in support of this research. **Author contributions:** The authors contributed equally to all aspects of this work and manuscript preparation. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data associated with this research may be found at www.cs.dartmouth.edu/farid/downloads/publications/scienceadvances17. Additional data related to this paper may be requested from the authors.

Submitted 2 August 2017
Accepted 11 December 2017
Published 17 January 2018
10.1126/sciadv.aao5580

Citation: J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).

The accuracy, fairness, and limits of predicting recidivism

Julia Dressel and Hany Farid

Sci Adv **4** (1), eaao5580.
DOI: 10.1126/sciadv.aao5580