DRAFT IN PROGRESS

A simplified numerical example

Suppose we can give convicted criminals a simple test, that allows us to categorize them into one of two groups: low risk or high risk. We know from experience that those classified as “high risk” recidivate at a rate of 60%, and those classified as “low risk” at a rate of 30%. Imagine that we test 200 convicts, and find that exactly 100 are classified as “high” and 100 are classified as “low.” We would expect to see a total of 90 recidivators from these 200 convicts, with 60 from the “high” group and 30 from the “low” group.

Now imagine our prison is full, and we have to release one prisoner. Should we consider the results of the test? This is a difficult question, and some will object to using the test on principle. But remember, given our assumptions, releasing a “high” prisoner instead of “low” will on average result in an extra 0.3 recidivism events. These additional crimes have real costs, and the burdens probably fall disproportionately on the least fortunate in society. So it is not crazy to think we might want to use the test. From now on I will leave aside the question of legitimacy, and just focus on the question of bias.

As stipulated, our predictor is unbiased: we predict that 60% of the “high” group and 30% of the “low” group will recidivate, and that is indeed what happens.

Now let’s add race. Assume there are 100 blacks and 100 whites in our convict pool. Assume further that the test-based predictor is unbiased by race: in other words, exactly 60% of blacks classified as “high risk” recidivate, as do 60% of whites so classified. Also, exactly 30% of blacks classified as “low risk” recidivate, as do 30% of whites so classified.

But let’s further assume that the true level of recidivism differs by race (as it does in the real world data). Because our test is unbiased, this also means that more blacks than whites will be classified as high risk. This is not an example of bias, it is a natural result of higher recidivism among blacks.

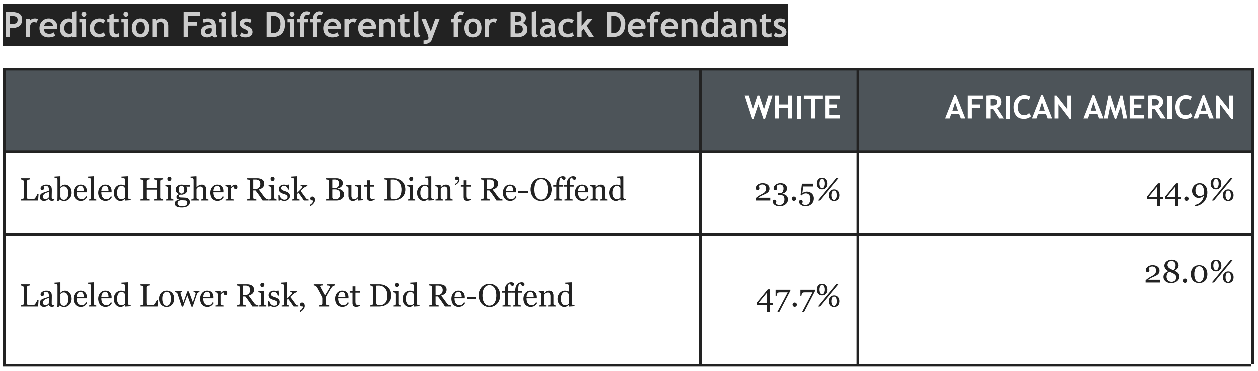
To put numbers on it, assume that the “high risk” group contains 60 blacks and 40 whites, while the “low risk” group contains 40 blacks and 60 whites. Then we would expect the following outcomes for our 200 convicts:

	White		Black		All	
	Lo	Hi	Lo	Hi	Lo	Hi
No Recid	42	16	28	24	70	40
Recid	18	24	12	36	30	60
Total	60	40	40	60	100	100

	False Pos	27.59%	False Pos	46.15%		
	False Neg	42.86%	False Neg	25.00%		

These results are fairly close to those described in the ProPublica result (also see the technical appendix linked from the main article). In particular, note that the “false positive” and “false negative” rates differ by race. This is NOT an indicator of bias, we assumed our test gave unbiased predictions for both races. Rather it is entirely a result of higher recidivism rates among blacks. Because there are more blacks classified as “high risk,” there are more black false positives and ALSO more black true positives.

From Pro Publica piece, very similar numbers:



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn’t Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

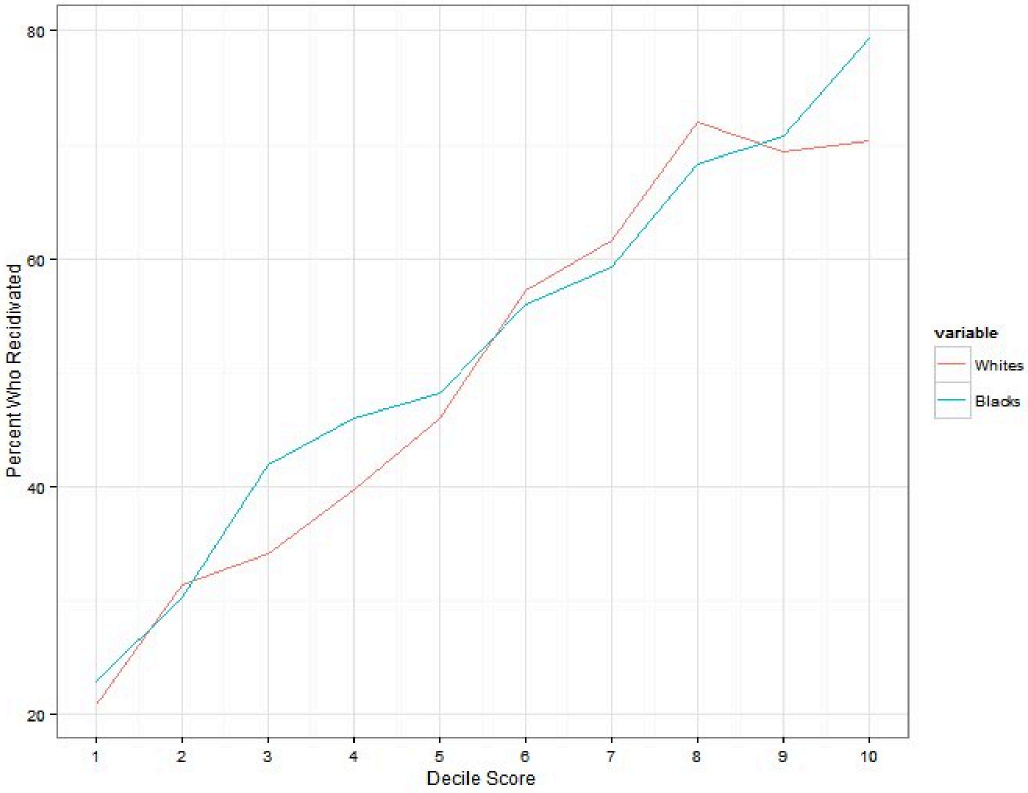
In short, the statistics that ProPublica believes are a smoking gun are nothing more than what you would expect to find if you have an unbiased predictor applied to two groups with different base rates of recidivism.

Better measure of bias
The previous numerical example showed how comparing “false positive” and “false negative” rates between groups is NOT an appropriate measure of bias. Bias would mean that blacks within a given classification are less likely to re-offend than whites with the same classification.

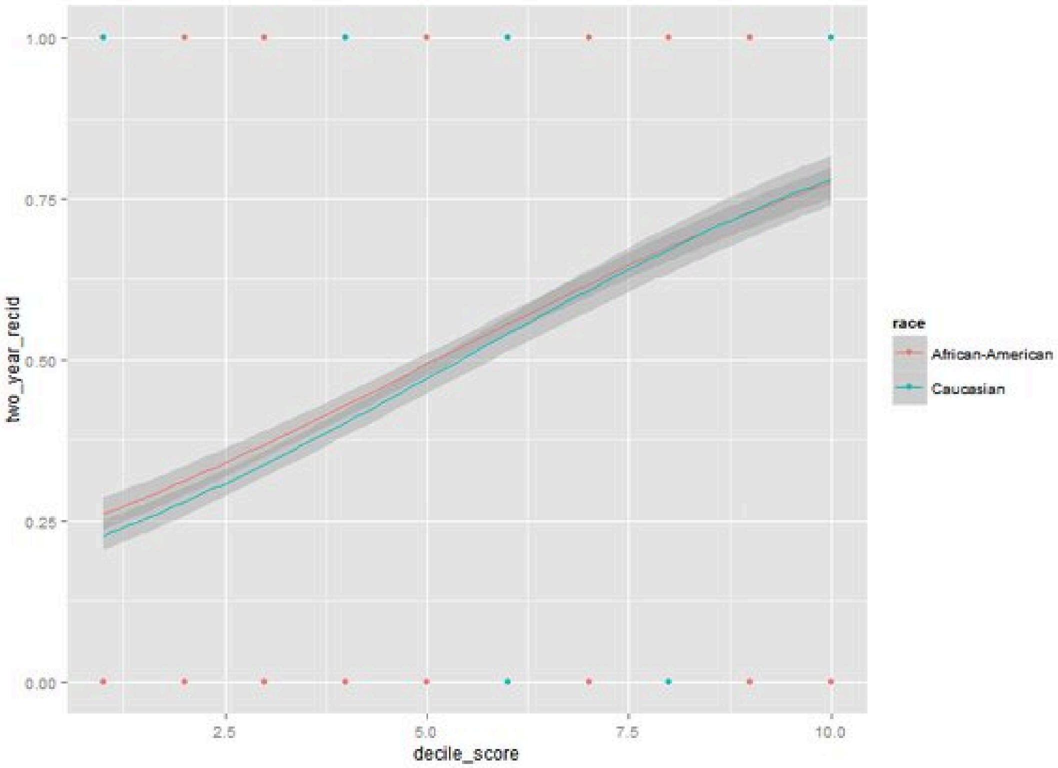
Lets look at the table of outcomes using the real data from ProPublica. We can immediately see that within both low and high risk categories, blacks are MORE likely to recidivate than are whites. On its face this would seem to be evidence that the algorithm is biased against whites, in favor of blacks.

However, this isn’t quite right either. The problems is that the “low” and “high” categories lump lots of people together. In reality, the algorithm produces something more like a continuous predictor. The reason “high” risk blacks is recidivate more is not that the predictor is biased against whites, but rather that the blacks in the high group are higher on the prediction continuum than whites are. So it would be more appropriate to look for bias in the more fine-grained, quasi-continuous predictor.

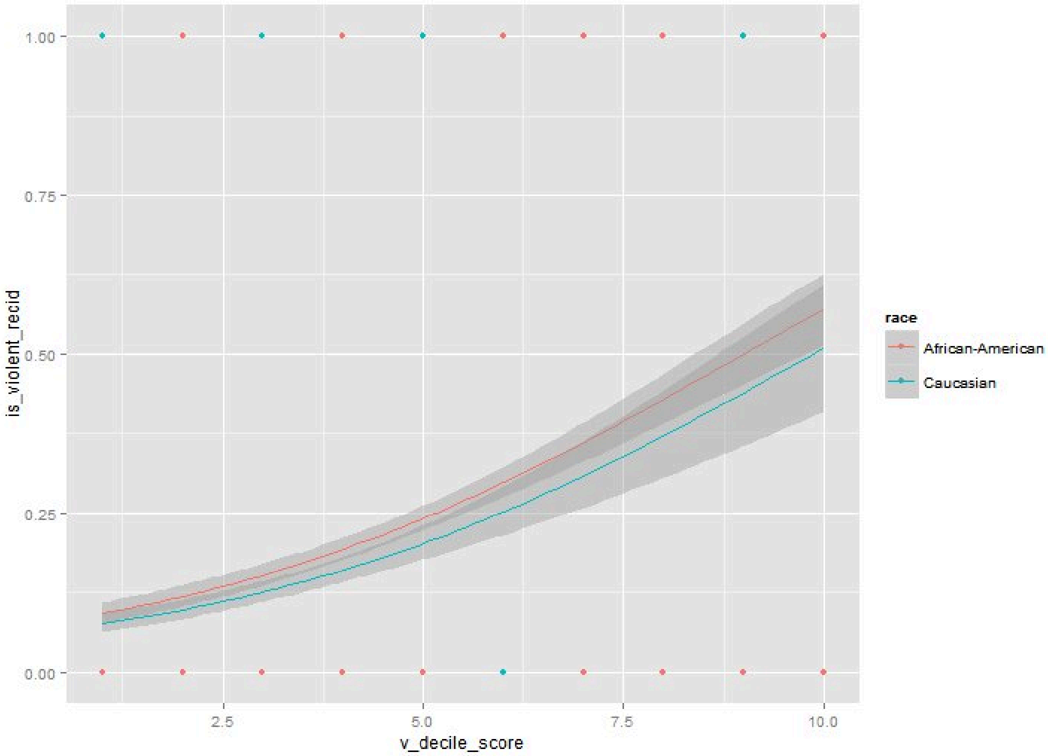
Robert VerBruggen, who writes, did this on the very day the ProPublica report came out. In a tweet he supplied the following figure showing rates of recidivism for each race as a function of the decile of risk classification. As you can see in the figure, there is absolutely no evidence in these data that the recidivism predictor is racially biased.



Smoothed version:



For violent crime:


is_violent_recid
v_decile_score
race
African-American
Caucasian