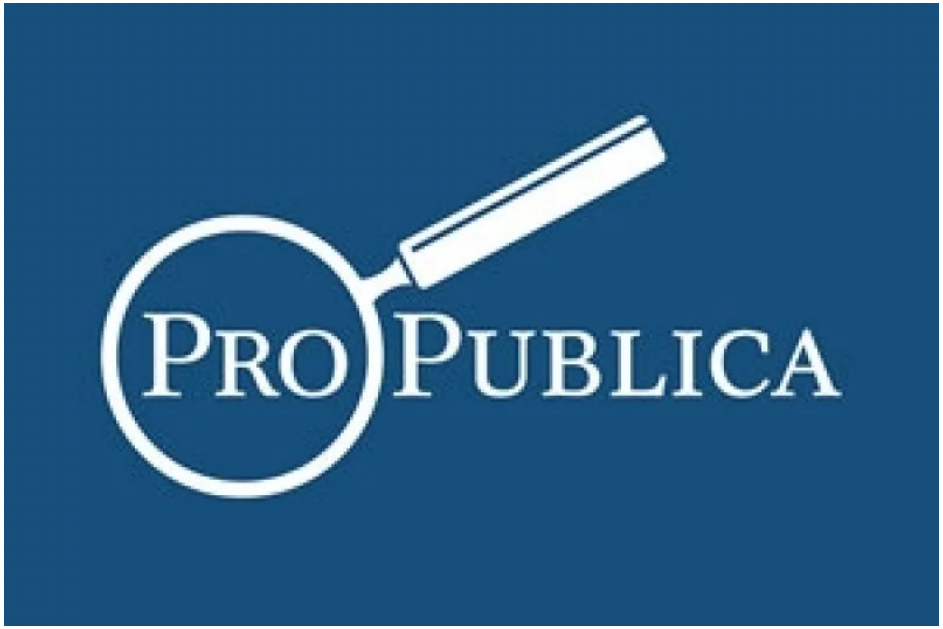
MACHINE BIAS

Technical Response to Northpointe

Northpointe asserts that a software program it sells that predicts the likelihood a person will commit future crimes is equally fair to black and white defendants. We re-examined the data, considered the company's criticisms, and stand by our conclusions.

by Jeff Larson and Julia Angwin, July 29, 2016, 11:55 a.m. EDT



On May 23, ProPublica published an article on a software program designed to predict the likelihood of future criminal conduct. The company that sells the program, Northpointe, responded with a 37-page critique of our work. We addressed the main thrust of the company's concerns here. Our replies to the company's methodological criticisms follow.

Northpointe allegation:

ProPublica neglected to consider the base rate in the interpretation of their results. This is an error in judgment about the probability of an event. The error occurs when information about the base rate of an event (e.g., low base rate of recidivism in a population) is ignored or not given enough weight.

ProPublica response:

This is not correct. ProPublica accounted for the difference in recidivism rates in a statistical test known as a logistic regression. That test found that when adjusting for recidivism, criminal history, age and gender across races, black defendants were 45 percent more likely to get a higher score. In addition, we calculated likelihood ratios, which are useful for assessing how well a test performs independent of base rate. The likelihood ratios we calculated showed that the Northpointe test performs differently across races. For black defendants, the likelihood ratio is lower than for white defendants. This means that a white defendant who has a higher score is more likely to recidivate than a black defendant who gets a higher score.

Northpointe allegation:

ProPublica combined the High and Medium levels and refer to this level in their article as "Higher Risk." Thus, PP's analysis of classification errors is for the Low cut point. This has the effect of inflating the false positive rate and the corresponding base-rate sensitive Target Population Error (1-PV+).

ProPublica response:

This is not correct. In our analysis, the disparity in false positive rates was starker when we adjusted the cut points as suggested by Northpointe. When adjusting the cut points to only include "High" risk scores, black defendants who did not go on to commit new crimes were three times as likely as white defendants to be classified as high risk.

ProPublica based its decision to group the medium and high scores together on Northpointe's user guide, which states: "scores in the medium

and high range garner more interest from supervision agencies."

Northpointe allegation:

ProPublica misrepresented the Model Errors as if they were Target Population Errors.

ProPublica response:

Not so. Northpointe is describing the positive and negative predictive values of the test. We reported the differences in true and false positive rates for black and white defendants.

Northpointe allegation:

ProPublica failed to report that the comparisons of the false positive rate and true positive rate for blacks and whites at the study cut point ("Low", "Not Low") for the VRRS and GRRS were inconclusive.

ProPublica response:

We did not report false positive rate and true positive rate comparisons because they do not offer additional insight into the question at hand: Are higher scores distributed unequally between black and white defendants? In fact they are. Black defendants who did not recidivate were more likely to be labeled higher risk.

Northpointe allegation:

The reverse logistic regression models are misspecified. And the relative risk ratios from the reverse regressions are miscalculated and misinterpreted.

ProPublica Response:

Our logistic model wasn't trying to predict who would recidivate. We were trying to identify a possible relationship between race and receiving a high score when controlling for other variables like age, gender and criminal history. We found that black defendants have greater odds of getting a high

score that cannot be explained by these other factors. Then we even controlled for future recidivism, and still found that the racial gap couldn't be explained.

Elsewhere in our methodology, we ask another question. Is the difference in recidivism risk between a high- and low-scoring black defendant different than that gap for a high- and low-scoring white defendant? If there is a difference in the increased risk associated with those scores within races that would mean a high score means something different for a black defendant than a white defendant.

In other words, when you compare black and white defendants with similar characteristics, black defendants tend to get higher scores.

Northpointe compared across races without correcting for other factors. But when making a comparison like this, it's necessary to correct for other factors like age, gender and prior crimes. When we did that, black defendants with higher scores were less risky than comparable white defendants. In other words, as underlying risk increases, scores will increase more for black defendants than white defendants. Which bolsters our finding in the logistic regression instead of contradicting it.

Northpointe allegation:

ProPublica conducted analyses in different samples that yield disparate results. The best AUC results were obtained in Sample A. Sample A consists of persons with complete case records.

ProPublica response:

This assertion is correct but misleading. ProPublica used two different samples in its analysis. In the logistic regression, which does not factor in time, we included only people for whom we could obtain two years' worth of recidivism data for an apples to apples comparison. The other analysis, known as a Cox regression, is able to take time into account. So for that analysis we could justifiably include cases where we did not have a full two-year window (i.e. those with less time to recidivate). This model did

show lower accuracy, as Northpointe points out, but that's not the result of unfairly manipulating the data, but rather using the most complete data possible compatible with the technique in use. One way of visualizing the difference in errors across scores such as those produced by Northpointe's risk assessment tool is something called an ROC curve. The curve visualizes the predictive power of a model. The more the curve bows toward the upper left-hand corner, the more accurate the test.

In its analysis, Northpointe presents ROC curves for black and white defendants, claiming that because the curves are very similar their visualization disproves ProPublica's analysis. However, the curves they included were "smoothed," a technique that concealed differences between black and white ROC curves. The problem with this is that a smoothed curve is appropriate for continuous variables, but the decile scores produced by Northpointe's tool are discrete (they must be a whole number between 1 and 10). This smoothing minimized the error differences between populations across scores. ProPublica plotted the unsmoothed curves here, clearly showing differences between the black and white ROC curves.

Northpointe allegation:

ProPublica misdefined the c-index as percent accuracy.

ProPublica response:

This is not accurate. ProPublica's detailed methodology paper never uses the term "percent accuracy." Instead we report concordance index values as an indicator of predictive accuracy. This is not a novel interpretation. In fact, in its own validation study published in 2008, Northpointe refers to the c-index as being a measure of "predictive accuracy."

Northpointe allegation:

There are overlapping time intervals in the Cox survival analysis data frame (Sample C). The stop-start time intervals in the survival data frame should not overlap. For example if the first start-stop time interval for a

case is 0–100, the next time interval should start after 100, but not before 100.

ProPublica response:

This is correct, but doesn't change the results. We analyzed 10,985 defendants who were assigned a risk score for Violent Recidivism, of which four have overlapping time intervals. We regret the errors, but excluding those cases would not change the outcome of our calculations.

Northpointe allegation:

Different norm sets may have been used for the decile scores. PP did not control for norm set. This would affect the location of the cut point for the study classifier ("Low" vs. "Not Low").

ProPublica response:

ProPublica used the designations of 'low', 'medium' and 'high' that Northpointe actually assigned to each of the defendants in our data. Following the advice in Northpointe's user manual, we collapsed the medium and high categories.

That manual states: "scores in the medium and high range garner more interest from supervision agencies than low scores, as a low score would suggest there is little risk of general recidivism," so we considered scores any higher than "low" to indicate a risk of recidivism in our analysis.

Northpointe allegation:

ProPublica describes the sample as pretrial defendants. It is not clear what the legal status was at the time of assessment for the cases in the sample.

ProPublica response:

Incorrect. The Broward County data used in our analysis indicates the legal status of the defendant at the time they were scored. ProPublica limited its analysis to defendants whose scores were assigned during pretrial because that is the primary use of the score in Broward County.

Jeff Larson

Jeff Larson is a reporter at ProPublica.

@thejefflarson

Julia Angwin

Julia Angwin is a senior reporter at ProPublica. From 2000 to 2013, she was a reporter at The Wall Street Journal, where she led a privacy investigative team that was a finalist for a Pulitzer Prize in Explanatory Reporting in 2011 and won a Gerald Loeb Award in 2010.

Julia.Angwin@propublica.org @JuliaAngwin