# How We Analyzed the COMPAS Recidivism Algorithm

*by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin*

May 23, 2016

← Read the story

Across the nation, judges, probation and parole officers are increasingly using algorithms to assess a criminal defendant's likelihood of becoming a recidivist – a term used to describe criminals who re-offend. There are dozens of these risk assessment algorithms in use. Many states have built their own assessments, and several academics have written tools. There are also two leading nationwide tools offered by commercial vendors.

We set out to assess one of the commercial tools made by Northpointe, Inc. to discover the underlying accuracy of their recidivism algorithm and to test whether the algorithm was biased against certain groups.

Our analysis of Northpointe's tool, called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.

We looked at more than 10,000 criminal defendants in Broward County, Florida, and compared their predicted recidivism rates with the rate that actually occurred over a two-year period. When most defendants are booked in jail, they respond to a COMPAS questionnaire. Their answers are fed into the COMPAS software to generate several scores including predictions of "Risk of Recidivism" and "Risk of Violent Recidivism."

We compared the recidivism risk categories predicted by the COMPAS tool to the actual recidivism rates of defendants in the two years after they were scored, and found that the score correctly predicted an offender's recidivism 61 percent of the time, but was only correct in its predictions of violent recidivism 20 percent of the time.

In forecasting who would re-offend, the algorithm correctly predicted recidivism for black and white defendants at roughly the same rate (59 percent for white defendants, and 63 percent for black defendants) but made mistakes in very different ways. It misclassifies the white and black defendants differently when examined over a two-year follow-up period.

Our analysis found that:

- Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Our analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).
- White defendants were often predicted to be less risky than they were. Our analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-

offenders (48 percent vs. 28 percent).

- The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.
- Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. And white violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists.
- The violent recidivism analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 77 percent more likely to be assigned higher risk scores than white defendants.

## Previous Work

In 2013, researchers Sarah Desmarais and Jay Singh examined 19 different recidivism risk methodologies being used in the United States and found that "in most cases, validity had only been examined in one or two studies conducted in the United States, and frequently, those investigations were completed by the same people who developed the instrument."

Their analysis of the research published before March2013 found that the tools "were moderate at best in terms of predictive validity," Desmarais said in an interview. And she could not find any substantial set of studies conducted in the United States that examined whether risk scores were racially biased. "The data do not exist," she said.

The largest examination of racial bias in U.S. risk assessment algorithms since then is a 2016 paper by Jennifer Skeem at University of California, Berkeley and Christopher T. Lowenkamp from the Administrative Office of the U.S. Courts. They examined data about 34,000 federal offenders to test the predictive validity of the Post Conviction Risk Assessment tool that was developed by the federal courts to help probation and parole officers determine the level of supervision required for an inmate upon release.

The authors found that the average risk score for black offenders was higher than for white offenders, but that concluded the differences were not attributable to bias.

A 2013 study analyzed the predictive validity among various races for another score called the Level of Service Inventory, one of the most popular commercial risk scores from Multi-Health Systems. That study found that "ethnic minorities have higher LS scores than nonminorities." The study authors, who are Canadian, noted that racial disparities were more consistently found in the U.S. than in Canada. "One possibility may be that systematic bias within the justice system may distort the measurement of 'true' recidivism," they wrote.

A smaller 2006 study of 532 male residents of a work-release program also found "a tendency toward classification errors for African Americans" in the Level of Service Inventory-Revised. The study, by Kevin Whiteacre of the Salvation Army Correctional Services Program, found that 42.7 percent of African Americans were incorrectly classified as high risk, compared with 27.7 percent of Caucasians and 25 percent of Hispanics. That study urged correctional facilities to investigate the their use of the scores independently using a simple contingency table approach that we follow later in this study.

As risk scores move further into the mainstream of the criminal justice system, policy makers have called for further studies of whether the scores are biased.

When he was U.S. Attorney General, Eric Holder asked the U.S. Sentencing Commission to study potential bias in the tests used at sentencing. "Although these measures were

crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society." The sentencing commission says it is not currently conducting an analysis of bias in risk assessments.

So ProPublica did its own analysis.

## How We Acquired the Data

We chose to examine the COMPAS algorithm because it is one of the most popular scores used nationwide and is increasingly being used in pretrial and sentencing, the so-called "front-end" of the criminal justice system. We chose Broward County because it is a large jurisdiction using the COMPAS tool in pretrial release decisions and Florida has strong open-records laws.

Through a public records request, ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida. We received data for all 18,610 people who were scored in 2013 and 2014.

Because Broward County primarily uses the score to determine whether to release or detain a defendant before his or her trial, we discarded scores that were assessed at parole, probation or other stages in the criminal justice system. That left us with 11,757 people who were assessed at the pretrial stage.

Each pretrial defendant received at least three COMPAS scores: "Risk of Recidivism," "Risk of Violence" and "Risk of Failure to Appear."

COMPAS scores for each defendant ranged from 1 to 10, with ten being the highest risk. Scores 1 to 4 were labeled by COMPAS as "Low"; 5 to 7 were labeled "Medium"; and 8 to 10 were labeled "High."

Starting with the database of COMPAS scores, we built a profile of each person's criminal history, both before and after they were scored. We collected public criminal records from the Broward County Clerk's Office website through April 1, 2016. On average, defendants in our dataset were not incarcerated for 622.87 days (sd: 329.19).

We matched the criminal records to the COMPAS records using a person's first and last names and date of birth. This is the same technique used in the Broward County COMPAS validation study conducted by researchers at Florida State University in 2010. We downloaded around 80,000 criminal records from the Broward County Clerk's Office website.

To determine race, we used the race classifications used by the Broward County Sheriff's Office, which identifies defendants as black, white, Hispanic, Asian and Native American. In 343 cases, the race was marked as Other.

We also compiled each person's record of incarceration. We received jail records from the Broward County Sheriff's Office from January 2013 to April 2016, and we downloaded public incarceration records from the Florida Department of Corrections website.

We found that sometimes people's names or dates of birth were incorrectly entered in some records – which led to incorrect matches between an individual's COMPAS score and his or her criminal records. We attempted to determine how many records were affected. In a random sample of 400 cases, we found an error rate of 3.75 percent (CI: +/- 1.8 percent).

## How We Defined Recidivism

Defining recidivism was key to our analysis.

In a 2009 study examining the predictive power of its COMPAS score, Northpointe defined recidivism as "a finger-printable arrest involving a charge and a filing for any uniform crime reporting (UCR) code." We interpreted that to mean a criminal offense that resulted in a jail booking and took place after the crime for which the person was COMPAS scored.

It was not always clear, however, which criminal case was associated with an individual's COMPAS score. To match COMPAS scores with accompanying cases, we considered cases with arrest dates or charge dates within 30 days of a COMPAS assessment being conducted. In some instances, we could not find any corresponding charges to COMPAS scores. We removed those cases from our analysis.

Next, we sought to determine if a person had been charged with a new crime subsequent to crime for which they were COMPAS screened. We did not count traffic tickets and some municipal ordinance violations as recidivism. We did not count as recidivists people who were arrested for failing to appear at their court hearings, or people who were later charged with a crime that occurred prior to their COMPAS screening.

For violent recidivism, we used the FBI's definition of violent crime, a category that includes murder, manslaughter, forcible rape, robbery and aggravated assault.

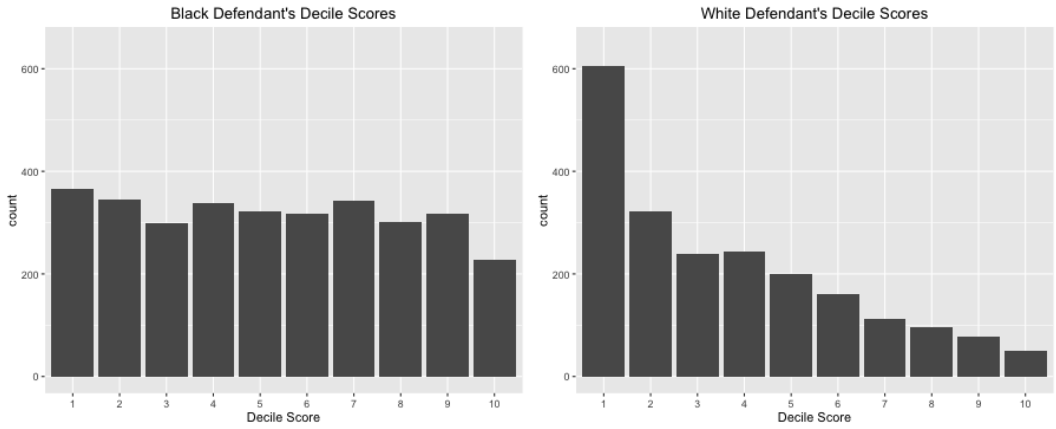
For most of our analysis, we defined recidivism as a new arrest within two years. We based this decision on Northpointe's practitioners guide, which says that its recidivism score is meant to predict "a new misdemeanor or felony offense within two years of the COMPAS administration date."

In addition, a recent study of 25,000 federal prisoners' recidivism rates by the U.S. Sentencing Commission, which shows that most recidivists commit a new crime within the first two years after release (if they are going to commit a crime at all).

## Analysis

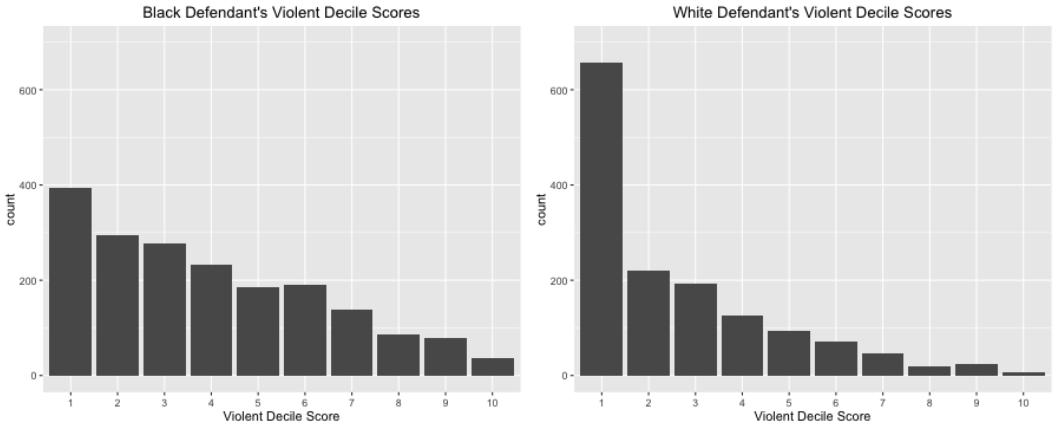
We analyzed the COMPAS scores for "Risk of Recidivism" and "Risk of Violent Recidivism." We did not analyze the COMPAS score for "Risk of Failure to Appear."

We began by looking at the risk of recidivism score. Our initial analysis looked at the simple distribution of the COMPAS decile scores among whites and blacks. We plotted the distribution of these scores for 6,172 defendants who had not been arrested for a new offense or who had recidivated within two years.



These histograms show that scores for white defendants were skewed toward lower-risk categories, while black defendants were evenly distributed across scores. In our two-year sample, there were 3,175 black defendants and 2,103 white defendants, with 1,175 female defendants and 4,997 male defendants. There were 2,809 defendants who recidivated within two years in this sample.

The histograms for COMPAS's violent risk score also show a disparity in score distribution between white and black defendants. The sample we used to test COMPAS's violent recidivism score was slightly smaller than for the general recidivism score: 4,020 defendants, 1,918 black defendants and 1,459 white defendants. There were 652 violent recidivists.




While there is a clear difference between the distributions of COMPAS scores for white and black defendants, merely looking at the distributions does not account for other demographic and behavioral factors.

To test racial disparities in the score controlling for other factors, we created a logistic regression model that considered race, age, criminal history, future recidivism, charge degree, gender and age.

**Risk of General Recidivism Logistic Model**

| *Dependent variable:* | |
|---|---|
| | Score (Low vs Medium and High) |
| Female | 0.221*** (0.080) |
| Age: Greater than 45 | -1.356*** (0.099) |
| Age: Less than 25 | 1.308*** (0.076) |
| Black | 0.477*** (0.069) |
| Asian | -0.254 (0.478) |
| Hispanic | -0.428*** (0.128) |
| Native American | 1.394* (0.766) |
| Other | -0.826*** (0.162) |
| Number of Priors | 0.269*** (0.011) |
| Misdemeanor | -0.311*** (0.067) |
| Two year Recidivism | 0.686*** (0.064) |
| Constant | -1.526*** (0.079) |
| Observations | 6,172 |
| Akaike Inf. Crit. | 6,192.402 |

*Note: *p<0.1; **p<0.05; ***p<0.01*

We used those factors to model the odds of getting a higher COMPAS score. According to Northpointe's practitioners guide, COMPAS "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.

Our logistic model found that the most predictive factor of a higher risk score was age. Defendants younger than 25 years old were 2.5 times as likely to get a higher score than middle aged offenders, even when controlling for prior crimes, future criminality, race and gender.

Race was also quite predictive of a higher score. While Black defendants had higher recidivism rates overall, when adjusted for this difference and other factors, they were 45 percent more likely to get a

higher score than whites.

Surprisingly, given their lower levels of criminality overall, female defendants were 19.4 percent more likely to get a higher score than men, controlling for the same factors.

**Risk of Violent Recidivism Logistic Model**

| | *Dependent variable:* |
|---|---|
| | Score (Low vs Medium and High) |
| Female | -0.729*** (0.127) |
| Age: Greater than 45 | -1.742*** (0.184) |
| Age: Less than 25 | 3.146*** (0.115) |
| Black | 0.659*** (0.108) |
| Asian | -0.985 (0.705) |
| Hispanic | -0.064 (0.191) |
| Native American | 0.448 (1.035) |
| Other | -0.205 (0.225) |
| Number of Priors | 0.138*** (0.012) |
| Misdemeanor | -0.164* (0.098) |
| Two Year Recidivism | 0.934*** (0.115) |
| Constant | -2.243*** (0.113) |
| Observations | 4,020 |
| Akaike Inf. Crit. | 3,022.779 |

*Note: *p<0.1; **p<0.05; ***p<0.01*

The COMPAS software also has a score for risk of violent recidivism. We analyzed 4,020 people who were scored for violent recidivism over a period of two years (not including time spent incarcerated). We ran a similar regression model for these scores.

Age was an even stronger predictor of a higher score for violent recidivism. Our regression showed that young defendants were 6.4 times more likely to get a higher score than middle age defendants, when correcting for criminal history, gender, race and future violent recidivism.

Race was also predictive of a higher score for violent recidivism. Black defendants were 77.3 percent more likely than white defendants to receive a higher score, correcting for criminal history and future violent recidivism.

---

To test COMPAS's overall predictive accuracy, we fit a Cox proportional hazards model to the data – the same technique that Northpointe used in its own validation study. A Cox model allows us to compare rates of recidivism while controlling for time. Because we aren't controlling for other factors such as a defendant's criminality we can include more people in this Cox model. For this analysis our sample size was 10,314 defendants (3,569 white defendants and 5,147 black defendants).

**Risk of General Recidivism Cox Model**

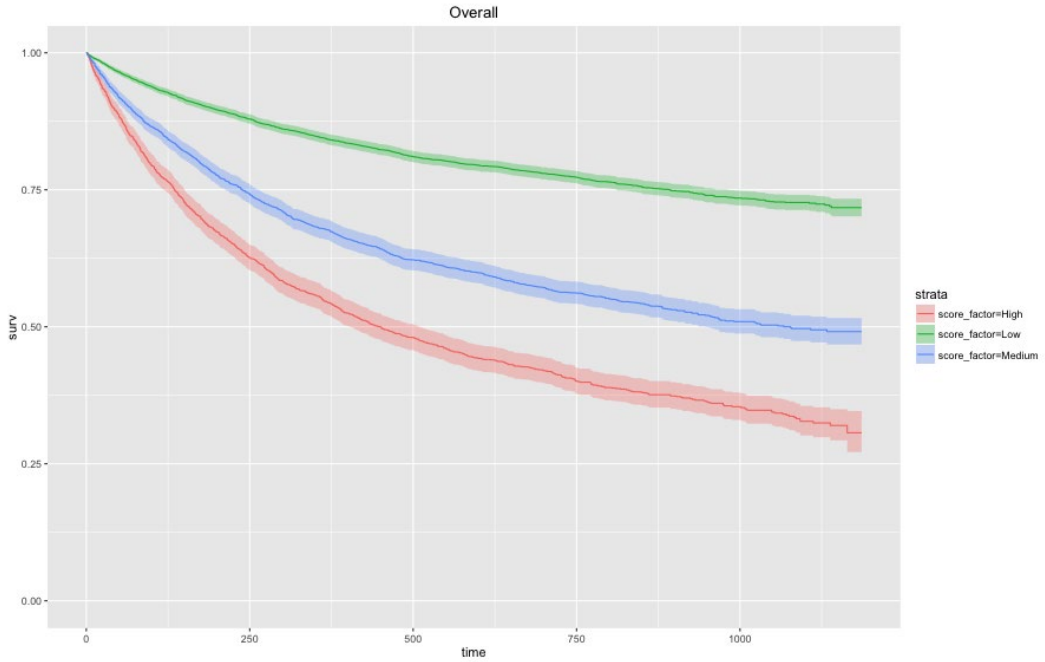
| | |
|---|---|
| High Risk | 1.250*** (0.041) |
| Medium Risk | 0.796*** (0.041) |
| Observations | 13,344 |
| R2 | 0.068 |
| Max. Possible R2 | 0.990 |
| Wald Test | 954.820*** (df = 2) |
| LR Test | 942.824*** (df = 2) |
| Score (Logrank) Test | 1,054.767*** (df = 2) |

*Note: *p<0.1; **p<0.05; ***p<0.01*

We considered people in our data set to be "at risk" from the day they were given the COMPAS score until the day they committed a new offense or April 1, 2016, whichever came first. We removed people from the risk set while they were incarcerated. The independent variable in the Cox model was the COMPAS categorical risk score.

The Cox model showed that people with high scores were 3.5 times as likely to recidivate as people in the low (scores 1 to 4) category. Northpointe's study, found that people with high scores (scores 8 to 10) were 5.6 times as likely to recidivate. Both results indicate that the score has predictive value.

A Kaplan Meier survival plot also shows a clear difference in recidivism rates between each COMPAS score level.



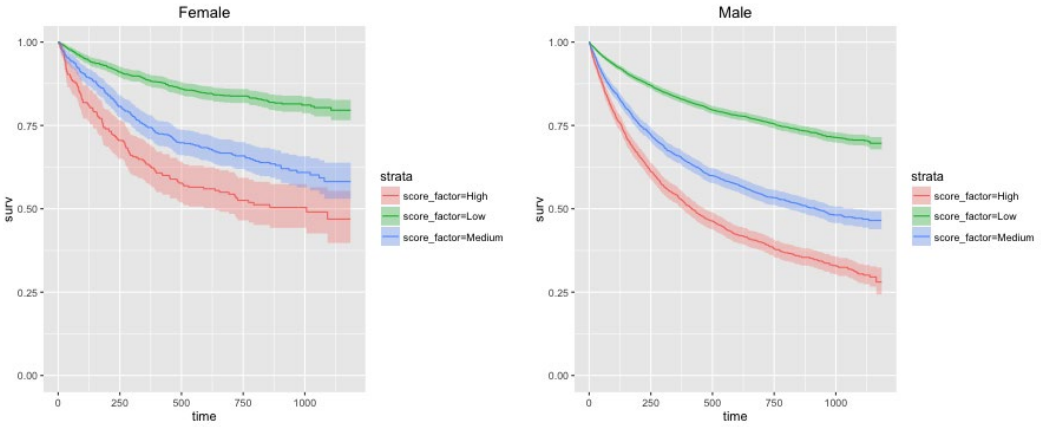

Overall, the Cox regression had a concordance score of 63.6 percent. That means for any randomly selected pair of defendants in the sample, the COMPAS system can accurately rank their recidivism risk 63.6 percent of the time (e.g. if one person of the pair recidivates, that pair will count as a successful match if that person also had a higher score). In its study, Northpointe reported a slightly higher concordance: 68 percent.

Running the Cox model on the underlying risk scores - ranked 1 to 10 - rather than the low, medium and high intervals yielded a slightly higher concordance of 66.4 percent.

Both results are lower than what Northpointe describes as a threshold for reliability. “A rule of thumb according to several recent articles is that AUCs of .70 or above typically indicate satisfactory predictive accuracy, and measures between .60 and .70 suggest low to moderate predictive accuracy,” the company says in its study.

The COMPAS violent recidivism score had a concordance of 65.1 percent.

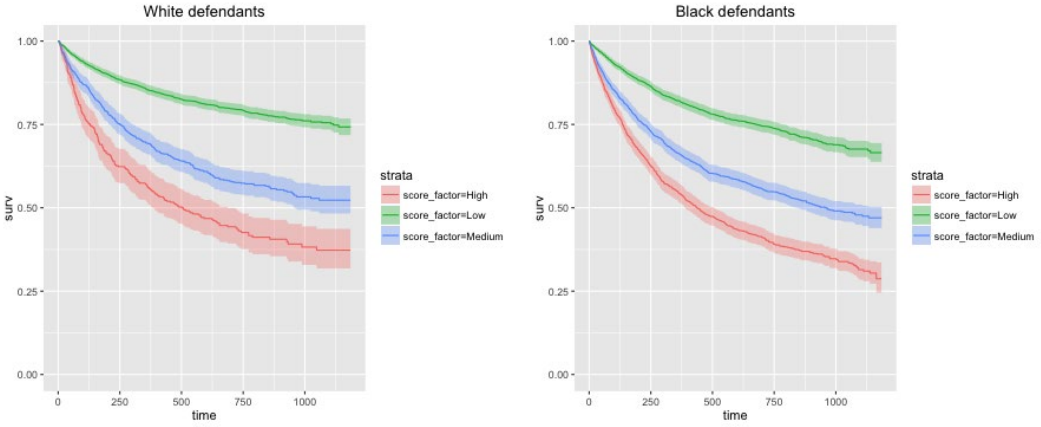
The COMPAS system unevenly predicts recidivism between genders. According to Kaplan-Meier estimates, women rated high risk recidivated at a 47.5 percent rate during two years after they were scored. But men rated high risk recidivated at a much higher rate – 61.2 percent – over the same time period. This means that a high-risk woman has a much lower risk of recidivating than a high-risk man, a fact that may be overlooked by law enforcement officials interpreting the score.




Northpointe does offer a custom test for women, but it is not in use in Broward County.

The predictive accuracy of the COMPAS recidivism score was consistent between races in our study – 62.5 percent for white defendants vs. 62.3 percent for black defendants. The authors of the Northpointe study found a small difference in the concordance scores by race: 69 percent for white defendants and 67 percent for black defendants.

Across every risk category, black defendants recidivated at higher rates.




| Risk of General Recidivism Cox Model (with Interaction Term) | |
|---|---|
| Black | 0.279*** (0.061) |
| Asian | -0.777 (0.502) |
| Hispanic | -0.064 (0.097) |
| Native American | -1.255 (1.001) |
| Other | 0.014 (0.110) |
| High Score | 1.284*** (0.084) |
| Medium Score | 0.843*** (0.071) |
| Black:High | -0.190* (.100, p: 0.0574) |
| Asian:High | 1.316* (0.768) |
| Hispanic:High | -0.119 (0.198) |
| Native American:High | 1.956* (.083) |
| Other:High | 0.415 (0.259) |

We also added a race-by-score interaction term to the Cox model. This term allowed us to consider whether the difference in recidivism between a high score and low score was different for black defendants and white defendants.

The coefficient on high scores for black defendants is almost statistically significant (0.0574). High-risk white defendants are 3.61 times as likely to recidivate as low-risk white defendants, while high-risk black defendants are only 2.99 times as likely to recidivate as low-risk black defendants. The hazard ratios for medium-risk defendants vs. low risk defendants also are different across races: 2.32 for white defendants and 1.95 for black defendants.

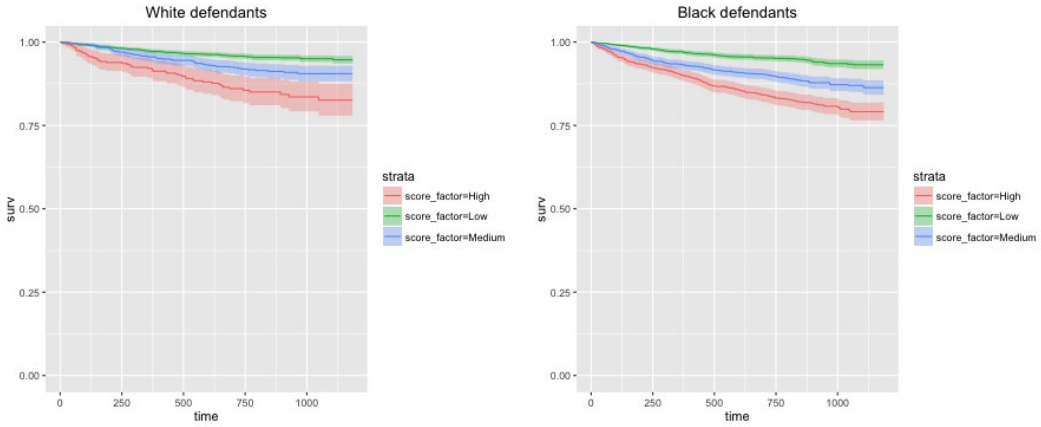
| | |
|---|---|
| Black:Medium | -0.173* (.091, p: 0.0578) |
| Asian:Medium | 0.986 (0.711) |
| Hispanic:Medium | 0.065 (0.164) |
| Native American:Medium | 1.390 (1.120) |
| Other:Medium | -0.334 (0.232) |
| Observations | 13,344 |
| R2 | 0.072 |
| Max. Possible R2 | 0.990 |
| Log Likelihood | -30,280.410 |
| Wald Test | 988.830*** (df = 17) |
| LR Test | 993.709*** (df = 17) |
| Score (Logrank) Test | 1,104.894*** (df = 17) |

*Note: *p<0.1; **p<0.05; ***p<0.01*

Because of the gap in hazard ratios, we can conclude that the score is performing differently among racial subgroups.

We ran a similar analysis on COMPAS's violent recidivism score, however we did not find a similar result. Here, we found that the interaction term on race and score was not significant, meaning that there is no significant difference the hazards of high and low risk black defendants and high and low risk white defendants.

Overall, there are far fewer violent recidivists than general recidivists and there isn't a clear difference in the hazard rates across score levels for black and white recidivists. These Kaplan Meier plots show very low rates of violent recidivism.




Finally, we investigated whether certain types of errors – false positives and false negatives – were unevenly distributed among races. We used contingency tables to determine those relative rates following the analysis outlined in the 2006 paper from the Salvation Army.

We removed people from our data set for whom we had less than two years of recidivism information. The remaining population was 7,214 – slightly larger than the sample in the logistic models above, because we don't need a defendant's case information for this analysis. As in the logistic regression analysis, we marked scores other than "low" as higher risk. The following tables show how the COMPAS recidivism score performed:

| **All Defendants** | | |
|---|---|---|
| | Low | High |
| Survived | 2681 | 1282 |
| Recidivated | 1216 | 2035 |
| FP rate: 32.35 | | |
| FN rate: 37.40 | | |
| PPV: 0.61 | | |
| NPV: 0.69 | | |
| LR+: 1.94 | | |
| LR-: 0.55 | | |

| **Black Defendants** | | |
|---|---|---|
| | Low | High |
| Survived | 990 | 805 |
| Recidivated | 532 | 1369 |
| FP rate: 44.85 | | |
| FN rate: 27.99 | | |
| PPV: 0.63 | | |
| NPV: 0.65 | | |
| LR+: 1.61 | | |
| LR-: 0.51 | | |

| **White Defendants** | | |
|---|---|---|
| | Low | High |
| Survived | 1139 | 349 |
| Recidivated | 461 | 505 |
| FP rate: 23.45 | | |
| FN rate: 47.72 | | |
| PPV: 0.59 | | |
| NPV: 0.71 | | |
| LR+: 2.23 | | |
| LR-: 0.62 | | |

These contingency tables reveal that the algorithm is more likely to misclassify a black defendant as higher risk than a white defendant. Black defendants who do not recidivate were nearly twice as likely to be classified by COMPAS as higher risk compared to their white counterparts (45 percent vs. 23 percent). However, black defendants who scored higher did recidivate slightly more often than white defendants (63 percent vs. 59 percent).

The test tended to make the opposite mistake with whites, meaning that it was more likely to wrongly predict that white people would not commit additional crimes if released compared to black defendants. COMPAS under-classified white reoffenders as low risk 70.5 percent more often than black reoffenders (48 percent vs. 28 percent). The likelihood ratio for white defendants was slightly higher 2.23 than for black defendants 1.61.

We also tested whether restricting our definition of high risk to include only COMPAS's high score, rather than including both medium and high scores, changed the results of our analysis. In that scenario, black defendants were three times as likely as white defendants to be falsely rated at high risk (16 percent vs. 5 percent).

We found similar results for the COMPAS violent recidivism score. As before, we calculated contingency tables based on how the score performed:

| All Defendants | | |
|---|---|---|
| | Low | High |
| Survived | 4121 | 1597 |
| Recidivated | 347 | 389 |
| FP rate: 27.93 | | |
| FN rate: 47.15 | | |
| PPV: 0.20 | | |
| NPV: 0.92 | | |
| LR+: 1.89 | | |
| LR-: 0.65 | | |

| Black defendants | | |
|---|---|---|
| | Low | High |
| Survived | 1692 | 1043 |
| Recidivated | 170 | 273 |
| FP rate: 38.14 | | |
| FN rate: 38.37 | | |
| PPV: 0.21 | | |
| NPV: 0.91 | | |
| LR+: 1.62 | | |
| LR-: 0.62 | | |

| White defendants | | |
|---|---|---|
| | Low | High |
| Survived | 1679 | 380 |
| Recidivated | 129 | 77 |
| FP rate: 18.46 | | |
| FN rate: 62.62 | | |
| PPV: 0.17 | | |
| NPV: 0.93 | | |
| LR+: 2.03 | | |
| LR-: 0.77 | | |

Black defendants were twice as likely as white defendants to be misclassified as a higher risk of violent recidivism, and white recidivists were misclassified as low risk 63.2 percent more often than black defendants. Black defendants who were classified as a higher risk of violent recidivism did recidivate at a slightly higher rate than white defendants (21 percent vs. 17 percent), and the likelihood ratio for white defendants was higher, 2.03, than for black defendants, 1.62.

*We've published the calculations and data for this analysis on github.*

← Read the story

---