COMPAS Risk Scales:
Demonstrating
Accuracy Equity and Predictive Parity

Performance
of the COMPAS Risk Scales
in Broward County

Northpointe Inc.
Research Department

William Dieterich, Ph.D.
Christina Mendoza, M.S.
Tim Brennan, Ph.D.

July 8, 2016

NORTHPOINTE

Contents

Chapter 1

Executive Summary

This research report presents Northpointe's technical analysis of the Broward County data that was used in ProPublica's article "Machine Bias" published by ProPublica on May 23, 2016 (Angwin, Larson, Mattu, & Kirchner, 2016).

We carefully examined the statistical methods in the article and conducted our own thorough analysis of the data that the ProPublica authors used for their study and made available through their web site.

Based on our examination of the work of Angwin et al. and on results of our analysis of their data, we strongly reject the conclusion that the COMPAS risk scales are racially biased against blacks. This report presents evidence that refutes the claim that the COMPAS risk scales were biased against black defendants in a sample of pretrial defendants in Broward County, Florida.

Our review leads us to believe that ProPublica made several statistical and technical errors such as misspecified regression models, wrongly defined classification terms and measures of discrimination, and the incorrect interpretation and use of model errors.

- ProPublica focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites. Their use of these statistics resulted in false assertions in their article that were repeated subsequently in interviews and in articles in the national media.
- When the correct classification statistics are used, the data do not substantiate the ProPublica claim of racial bias towards blacks.
- The proper interpretation of the results in the samples used by ProPublica demonstrates that the General Recidivism Risk Scale (GRRS) and Violent Recidivism Risk Scale (VRRS) are equally accurate for blacks and whites.

Chapter 2

Introduction

We carefully reviewed the statistical methods in the ProPublica (PP) article and conducted our own thorough analysis of the Broward County data that they used for their study and made available through their web site.

Based on our careful review and on results of our analyses, we strongly reject their conclusion that the COMPAS risk scales are racially biased against blacks. We present evidence that thoroughly refutes PP's conclusion.

2.1 Northpointe's Main Findings

- Angwin et al. used the incorrect classification statistics to frame the COMPAS risk scales as biased against blacks. They compared the complements of *Sensitivity* and *Specificity* for blacks and whites. These are operating characteristics calculated separately on recidivists only and non-recidivists only. They should have used the complements of the predictive values that take into account the base rate of recidivism. In their main table, the PP authors misrepresented the *percentage of non-recidivists with a positive test* result ("Not Low" risk level) as the *percentage of persons with a positive test result that did not recidivate* ("Labeled Higher Risk, But Didn't Re-Offend").

- If the correct classification statistics are used, then the PP authors' claim of racial bias is not supported. In comparison with whites, a slightly lower percentage of blacks were "Labeled Higher Risk, But Didn't Re-Offend" (37% vs. 41%). In comparison with whites, a slightly higher percentage of blacks were "Labeled Lower Risk, Yet Did Re-Offend" (35% vs. 29%). Thus the claim of racial bias against blacks is refuted. The results demonstrate *predictive parity* for blacks and whites at the study cut point used by Angwin et al. ("Low" vs. "Not Low") and at alternative cut points ("Not High" vs. "High").

- AUC results in the PP study samples demonstrate that the General Recidivism Risk Scale (GRRS) and Violent Recidivism Risk Scale (VRRS) are equally accurate for blacks and whites (equal discriminative ability). Thus the risk scales exhibit *accuracy equity*. The AUC for the GRRS decile score predicting any arrest in PP's main sample is 0.69 (0.68, 0.71) for blacks and 0.69 (0.67, 0.71) for whites. The AUC in the overall sample is 0.70 (0.69, 0.71). The AUC for the VRRS decile score predicting any violent arrest in PP's main sample is 0.67 (0.64, 0.69) for blacks and 0.66 (0.62, 0.70) for whites. The AUC in the overall sample is 0.68 (0.66, 0.70).

Chapter 3

Results

3.1 ProPublica's Broward County Samples

In their article Angwin et al. present results from analyses conducted in three different samples. *Sample A* consists of pretrial defendants with complete case records who have at least two years of follow-up time. The PP authors use *Sample A* to fit reverse logistic regressions predicting the "Not Low" Risk Level. Subsets of *Sample A* are used for tests of the GRRS (n=6,172) and the VRRS (n=4,020). *Sample B* consists of pretrial defendants with at least two years of follow-up and possibly incomplete case records. *Sample B* is used by PP to calculate the classification contingency tables from which they derive the main study results for false positive rates (fpr) and false negative rates (fnr). The PP authors use subsets of *Sample B* for tests of the GRRS (n=7,214) and the VRRS (n=6,454). *Sample C* consists of 10,994 pretrial defendants with varying follow-up times (1 to 1,186 days) and possibly incomplete case records, that is used to fit Cox survival models. The study data are available on the ProPublica web site, along with the annotated code necessary to construct the study data frames and reproduce the analyses and verify the results presented in the ProPublica article.

When reporting study results it is standard practice to describe the study sample in terms of characteristics that are key to the interpretation of the findings. Differences in the risk scores for blacks and whites are at the heart of the PP authors' claim that the risk scales are biased against blacks. Unfortunately, they did not provide descriptive statistics for blacks and whites for any of the samples that they used. Descriptive statistics for criminal history would have helped to explain why the risk scale scores of whites were shifted so much lower than blacks. They did provide plots showing the number of defendants distributed across the decile score levels of the GRRS and VRRS for blacks and whites in the main sample (Sample B). Figure 3.1 displays bar plots of the percentage of defendants in the decile score levels of each risk scale by race. We reproduce the bar plots to point out that the decile scores of blacks have a better distribution in comparison with whites. The bar plots show that the GRRS decile scores of blacks are well-aligned with the norm group that is used in Broward County (close to 10% fall into each decile score level). The GRRS decile score distribution of white defendants on the other hand is shifted much lower relative to the norm. The distribution of the VRRS decile scores for black defendants is also better aligned with the norm group in comparison to white defendants. The distributions of whites are shifted lower because they have lower values on the inputs of the risk scales.

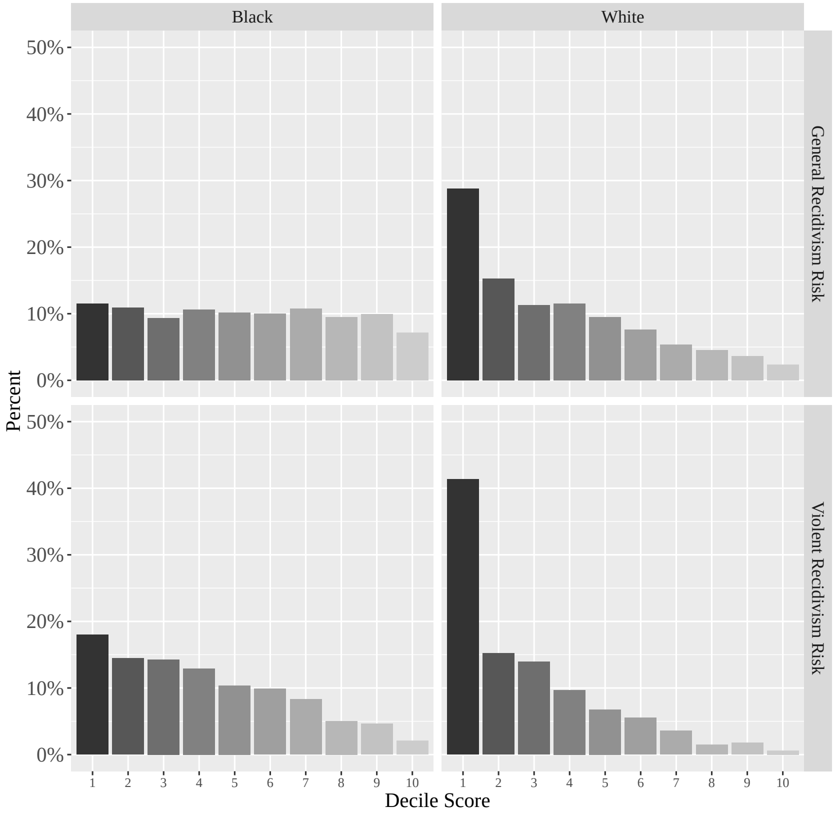


Figure 3.1: Percentage of defendants in the Decile Score Levels of Each Risk Scale by Race.

The inputs (risk factors) for the GRRS are the Criminal Involvement Scale, drug problems sub-scale, age at assessment, age at first adjudication, number of prior arrests, arrest rate, and the Vocational Educational Scale. The inputs (risk factors) for the VRRS are age at assessment, age at first adjudication, the History of Violence Scale, the History of Noncompliance Scale, and the

Vocational Educational Scale. The inputs for the GRRS and VRRS were not included in the study data sets that PP made available on their web site. PP did include age and number of prior arrests in the data they posted. We can compare blacks and whites in Sample B on these two factors. The white sample has less criminal history. The mean number of prior arrests is lower for whites (M=2.6, SD=3.8) compared with blacks (M=4.4, SD=5.6). The white sample is older on average. The mean age at assessment is higher for whites (M=37.7, SD=12.8) compared with blacks (M=32.7, SD=10.9).

We also point out that in comparison with blacks, whites have much lower base rates of general recidivism (0.39 vs. 0.51) and violent recidivism (0.09 vs. 0.14) in Sample B. The risk scores of whites should be shifted lower if the risk scales are valid predictors of recidivism.

3.2 Overview of Classification Statistics

In this section we provide an overview of classification statistics. The overview is necessary to understand the flaws in PP's analysis and interpretation of results. The material is unavoidably technical.

Types of Classification Statistics

Before being put into practice, a risk scale is often cut into levels, for example, Low, Medium, and High. This requires two thresholds that are then used for decision making. For instance, persons scoring above the High-Risk threshold are targeted for a more intensive treatment program. Making decisions at thresholds of the risk scale is a type of classification. A study is usually conducted to evaluate the intrinsic accuracy of the classifier and the performance of the classifier in practice.

There are two main types of classification statistics reported in the PP article: 1) *Model Errors* and 2) *Target Population Errors*. The PP article primarily focuses on *Model Errors* and presents these as evidence of racial bias. *Target Population Errors* are what should be analyzed if one is interested in testing for racial bias, but these are mostly ignored by the PP authors. To understand the mistakes that the PP authors made, it is necessary to briefly define these two types of classification errors. Full definitions of these and other classification terms are provided in appendix C.

Model Errors

The *Sensitivity* of the classifier is the percentage of recidivists correctly classified as recidivists. It is calculated on recidivists only.

The *Specificity* of the classifier is the percentage of non-recidivists correctly classified as non-recidivists. It is calculated on non-recidivists only.

The complement of *Sensitivity* is the *false negative rate* which is the percentage of recidivists misclassified as non-recidivists. The complement of *Specificity* is the *false positive rate* which is the percentage of non-recidivists misclassified as recidivists. These are the *Model Errors* that PP used as evidence of racial bias.

Sensitivity and *Specificity* quantify the accuracy of the risk scale. These classification statistics are useful for summarizing the accuracy of a risk scale. For instance they are used to describe the receiver operating characteristic (ROC) curve to estimate the area under the curve (AUC), one of the most widely used measures of diagnostic accuracy. The Receiver Operating Characteristic (ROC) method and the area under the ROC curve (AUC) are explained in the appendix C.

Angwin et al. could have properly used the operating characteristics as evidence of accuracy. But PP ignored and obfuscated the evidence that showed the AUCs obtained for the risk scales were the same, and thus equitable, for blacks and whites. A summary and discussion of our AUC results can be found in section 3.3.

Instead, the PP authors misused the operating characteristics as evidence of racial bias. They were wrong in doing that. *Model Errors* are of no practical use to a practitioner in a criminal justice agency who is assessing an offender's probability of recidivating. The practitioner does not know at the time of the assessment if the offender is a recidivist or not. *Model Errors* cannot be directly applied to an offender at the time of assessment (see Linn, 2004, for example).

It is unrealistic to expect equal *Model Error* trade-offs at a particular cut point in two samples that have different risk score distributions and base rates of recidivism. Although *Sensitivity* and *Specificity* do not depend on the base rate of recidivism, because they are calculated separately on recidivists and non-recidivists, the trade-offs between *Sensitivity* and *Specificity* can be impacted by the base rate. Leeflang, Rutjes, Reitsma, Hooft, and Bossuyt (2013), using data from 23 meta-analyses to assess the effect of base rate (prevalence) on *Sensitivity* and *Specificity*, found that *Specificity* decreased as disease base rate increased. In other words the *false positive rate* increased with increasing base rate. Differences in the base rates of blacks and whites for general recidivism

(0.51 vs. 0.39) and violent recidivism (0.14 vs. 0.09) in the PP samples strongly affected the *Sensitivity* and *Specificity* tradeoffs observed in the PP study.

We have conducted our own simulation analyses to assess the effects of differences in the risk scale distribution and base rate on the false positive and false negative rates. Results of our analyses indicate that as the mean difference in scores between a low-scoring group and a high-scoring group is increased, the base rates diverge and higher false positive rates and lower false negative rates are obtained for the high-scoring group. This is the same pattern of results reported by Angwin et al. This pattern does *not* show evidence of bias, but rather is a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores. These results help to explain the effects of the relatively higher risk scores and higher base rates of blacks on the false positive and false negative rates in the PP study. These results also make clear that it is not proper to make an assessment of racial bias on the basis of *Model Errors* obtained from the ROC method and detached from the base rate of deployment.

Target Population Errors

The whole purpose of administering a risk scale is to use the results to assess a person's risk of re-offending at the time of assessment in a particular agency. As discussed above the operating characteristics are not useful for this purpose. The useful classification statistic for this purpose is the predictive value. The *Positive Predictive Value* (PV+) is the probability that a person with a positive test result ("Not Low" risk score) will recidivate. The *Negative Predictive Value* (PV-) is the probability that a person with a negative test result ("Low" risk score) will not recidivate. *Sensitivity* and *Specificity* quantify the accuracy of the risk scale and the predictive value quantifies its clinical value (Pepe, 2003). A useful prediction will have a PV+ that is greater than the base rate and a PV- that is greater than 1 minus the base rate. A perfect test will predict the outcome perfectly with PV+ = 1 and PV- = 1. The predictive values depend on the accuracy of the test and the base rate of failure.

The complement of the PV+ (1 - PV+) is the probability of not recidivating given a positive test result ("Not Low" risk score). The complement of the PV- (1 - PV-) is the probability of recidivating given a negative test result ("Low" risk score). These are the *Target Population Errors* that should be examined if one is interested in determining if the risk scales perform differently for blacks and whites.

Impact of PP's Study Cut Point on the Classification Errors

Angwin et al. chose to conduct their analyses of classification errors at a cut point on the decile score that yields higher *Model Errors* and *Target Population Errors*. They combined the High and Medium levels and refer to this level in their article as "Higher Risk." Thus, their analysis of classification errors is for the Low cut point. This has the effect of inflating the false positive rate (fpr) and the corresponding base-rate-sensitive *Target Population Error* (1-PV+). They justify the decision to examine the Low threshold by pointing to a sentence in the *Practitioner's Guide to COMPAS Core* (Northpointe Inc., 2015b) which states that "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." More definitive guidance on the interpretation of decile scores is provided in *Measurement and Treatment Implications for COMPAS Core* which states that "Decile scores 1-4 may be regarded as Low Risk since they are clearly lower than "average." Decile Scores from 5-7 may be regarded as Medium Risk since they are in the middle of the distribution and represent cases that are very close to "average" for the total population of the agency. Decile Scores of 8 and above may be regarded as High Risk since they are in the top third of the distribution" (Northpointe Inc., 2015a). In this report we refer to the two levels examined by the PP authors as "Low" and "Not Low."

3.3 Accuracy Equity and Predictive Parity

A risk scale exhibits *accuracy equity* if it can discriminate recidivists and non-recidivists equally well for two different groups such as blacks and whites. The risk scale exhibits *predictive parity* if the classifier obtains similar predictive values for two different groups such as blacks and whites, for example, the probability of recidivating, given a high risk score, is similar for blacks and whites. The interpretation of relative predictive values is discussed in Appendix A.

Angwin et al. base their assessment of racial bias on an unrealistic criterion. Their requirement that the risk scale classification obtains the same *Sensitivity* and *Specificity* for blacks and whites at a particular cut point is unrealistic because the two groups have different risk scale distributions and different base rates. The PP authors ignore evidence of accuracy equity and predictive parity of the COMPAS risk scales.

AUC results in the PP samples demonstrate that the General and Violent recidivism risk scales are equally accurate for blacks and whites.[1] The AUC

[1] Accuracy refers to how accurately the risk scale discriminates between non-recidivists

results are in section 3.3. Thus the risk scales exhibit accuracy equity. The results indicate that blacks and whites obtain similar positive and negative predictive values using a classifier based on the study cut point ("Low" vs. "Not Low") and alternative cut points ("Not High" vs. "High"). Thus the risk scale classifiers exhibit predictive parity for blacks and whites.

Demonstrating the Predictive Parity of the COMPAS Risk Scales

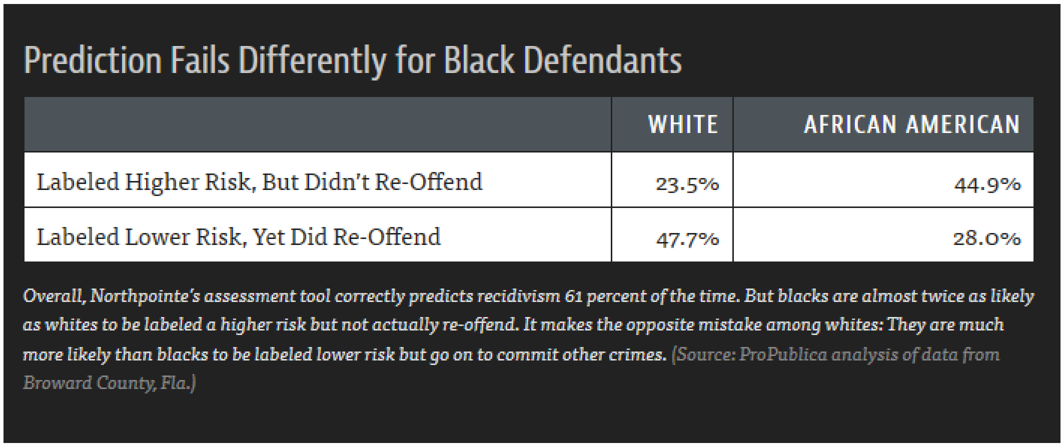
General Recidivism Risk Scale Classification Results

In their article Angwin et al. note that blacks have a much higher false positive rate (fpr) compared with whites and that whites have a much higher false negative rate (fnr) compared with blacks. They report that the fpr is 44.9% for blacks and 23.5% for whites. They also report that the fnr is 47.7% for whites and 28.0% for blacks. The PP authors interpretation of these classification errors is incorrect. The rates that they report are actually *Model Errors* that ignore the base rate of recidivism in the Broward County population. The PP authors misrepresents these *Model Errors* (operating characteristics) as if they are the *Target Population Errors* (predictive values) that would be obtained in the Broward County population using the base rates of recidivism in Broward County.

Figure 3.2 is a screenshot of the table of model errors that the PP authors presented under the caption "Prediction Fails Differently for Black Defendants." In the table, they incorrectly report the proportion of non-recidivists that have a "Not Low" risk score in the row named "Labeled Higher Risk, But Didn't Re-Offend" and incorrectly report the proportion of recidivists that have a "Low" risk score in the row named "Labeled Lower Risk, Yet Did Re-Offend." Thus the row names refer to the *Target Population Errors* that take into account the base rates of recidivism for blacks and whites, but the numbers are actually the *Model Errors* that are calculated separately for recidivists and non-recidivists and that ignore the base rates for blacks and whites. As discussed above, it is not appropriate to compare the *Model Errors* of blacks and whites.

Table 3.1 is a corrected version of the table that PP presented under the heading "Prediction Fails Differently for Black Defendants." We have added the correct *Target Population Errors* (predictive values) that take into account the base rates for blacks and whites. Recall that the complement of the *Positive Predictive Value* (1 - PV+) is the probability of not recidivating given a positive

and recidivists. The AUC is a rank-based measure of discriminative ability not a measure of accuracy.



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%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Figure 3.2: Screen shot of Propublica's table that incorrectly reports model errors at the study cut point (Low vs. Not Low) for the General Recidivism Risk Scale as if they are target population errors.

test result. The complement of the *Negative Predictive Value* (1 - PV-) is the probability of recidivating given a negative test result.

Now the numbers correctly correspond with the row names in PP's table. The results actually indicate that in comparison with whites a slightly lower percentage of blacks were "Labeled Higher Risk, But Didn't Re-Offend" (37% vs. 41%). The results also show that in comparison with whites, only a slightly higher percentage of blacks were "Labeled Lower Risk, Yet Did Re-Offend" (35% vs. 29%). Thus we conclude that the General Recidivism Risk Scale exhibits predictive parity for blacks and whites. This result refutes PP's claim of racial bias. Appendix A includes supplemental results including the classification statistics across all the decile scores.

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	41%	37%
Labeled Lower Risk, Yet Did Re-Offend	29%	35%

Table 3.1: Propublica's table with correct target population errors at the study cut point (Low vs. Not Low) for the General Recidivism Risk Scale.

Violent Recidivism Risk Scale Classification Results

Angwin et al. did not present a table of the *Model Errors* for the Violent Recidivism Risk Scale. But they did report the *Model Errors* and the predictive values for the Violent Recidivism Risk Scale. As we did with the General Recidivism Risk Scale classifier ("Low" vs. "Not Low"), we calculate the complements of the predictive values and table the correct *Target Population Errors*. Table 3.2 shows the *Target Population Errors* for the Violent Recidivism Risk Scale.

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	83%	79%
Labeled Lower Risk, Yet Did Re-Offend	7%	9%

Table 3.2: Target population errors at the study cut point (Low vs. Not Low) for the Violent Recidivism Risk Scale.

The results show that the percentage of blacks who were "Labeled Higher Risk, But Didn't Re-Offend" is slightly lower than the percentage of whites who were "Labeled Higher Risk, But Didn't Re-Offend" (79% vs. 83%). Note that these *Target Population Errors* are quite high. This is an example of the false positive paradox. This is a classification result obtained for a test applied to a low base rate outcome where the probability of a False Positive (1-PV+) is high even though the test is accurate. The result goes against our intuition that tells us the probability of a False Positive (1-PV+) should be lower when the base rate of recidivism is lower. But in practice, for a given risk scale, we find that the lower the base rate of recidivism in the population, the more likely it is that an offender predicted to recidivate will not recidivate.

The results in Table 3.2 also show that the percentage of blacks who were "Labeled Lower Risk, Yet Did Re-Offend" is only slightly higher than the percentage of whites who were "Labeled Lower Risk, Yet Did Re-Offend" (9% vs. 7%). Thus we conclude that the Violent Recidivism Risk Scale exhibits predictive parity for blacks and whites. This result refutes the PP authors' claim of racial bias.

Angwin et al. state in the main part of their article that the Violent Recidivism Risk Scale was "remarkably unreliable in forecasting violent crime." They note that "Only 20 percent of the people predicted to commit violent crimes actually went on to do so." They are referring to the positive predictive value. What they don't mention is that the base rate in the sample overall is only 11%. If the *Positive Predictive Value* is greater than the base rate, then the risk scale has clinical value. The *Positive Predictive Value* at the study cut point ("Low" vs. "Not Low") is 21% which is about twice the base rate. The PP authors are committing the base rate fallacy. This is an error in judgement 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 (Kahneman & Tversky, 1982). There is actually nothing remarkable about the positive predictive value pointed out by the PP authors. Appendix A includes the complete decision table across all the Violent Recidivism Risk Scale decile scores.

A more common version of the base rate fallacy was committed by the PP authors when they presented the *Model Errors* as if they were the *Target Population Errors* in their article. In fact, the way in which this fallacy is used by the authors to draw a controversial conclusion probably makes it one of the more stunning examples of the base rate fallacy in the literature.

Demonstrating the Accuracy Equity of the COMPAS Risk Scales

A risk scale exhibits accuracy equity if it can discriminate recidivists and non-recidivists equally well for two different groups such as blacks and whites. The most well-developed and widely used measure of discriminative ability is the area under the receiver operating characteristic (ROC) curve (AUC). The ROC curve is a plot of Sensitivity (tpr) and Specificity (tnr) for all possible cut points of a risk scale. The AUC is a summary measure of discriminative ability (diagnostic accuracy) across all the thresholds of the risk scale. The AUC is interpreted as the probability that a randomly selected recidivist will have a higher risk score than a randomly selected non-recidivist.

Angwin et al. downplayed results that indicated the AUCs were the same for blacks and whites. They focused on the operating characteristics at one point on the ROC curve. Their study cut point is a decile score greater than or equal to 5 (“Low” vs. “Not Low”). The PP authors misinterpreted these *Model Errors* (fnr, fpr) as if they were the *Target Population Errors* (1-PV-, 1-PV+). A proper interpretation of the operating characteristics obtained at the study cut point is provided in Appendix A.

The performance of the risk scale depends on its intrinsic accuracy as well as external factors such as base rate and user preferences about the costs of errors, benefits of treatment, selection ratio, and other criteria. The fact that the accuracy of the risk scale is the same for blacks and whites is critical. If the AUCs were not the same for blacks and whites, then the performance of the risk scales would not be similar for blacks and whites. Performance refers to the positive and negative predictive values and their complements that take into account the base rates for blacks and whites.

It appears that Angwin et al. selected certain combinations of methods and samples to obtain results that best supported their claims of racial bias. For example PP calculated the concordance index (*c*-index) for the three risk levels (Low, Medium, High) in the survival sample (*Sample C*) as opposed to calculating the area under the receiver operating characteristic curve (AUC) for the decile score in the classification contingency table sample (*Sample B*). The c-index is a generalization of the AUC used to measure the discriminative ability of a risk scale in survival data. The c-index is described in the glossary in Appendix C. The PP authors claim that they use the c-index because that measure of discrimination was used in a 2008 paper that examined the predictive validity of the COMPAS risk scales (Brennan, Dieterich, & Ehret, 2009).

It is hard to understand why Angwin et al. did not report the AUC for the decile score in *Sample B* in which they calculate the false negative rate (fnr) and false positive rate (fpr) that they use to support the central claim of their article. The most amplified result in their article was the finding of a higher fpr for blacks. The fpr was calculated in *Sample B*. The fpr and other model errors that they report are derived using receiver operating characteristic methods. Instead they report the c-index estimates in *Sample C* (survival sample). PP takes this approach despite the fact that the fpr and fnr cannot be calculated from the Cox survival model that they used. There is nothing wrong with using the c-index. In fact, for binary outcomes data without censoring such as used with logistic regression, the c-index and AUC are equivalent. In survival data with right censoring, the c-index is dependent on the pattern of censoring. The issues that are most concerning are the inconsistencies in the methods used and the samples analyzed.

An additional methods choice by the PP authors further biased results in favor of their claim of racial bias. Instead of reporting the AUC for the ten levels of the decile score, PP reports the c-index for the three risk levels (Low, Medium, and High). PP's c-index estimates for the risk levels (Low, Medium, and High) in *Sample C* are biased low in comparison to the AUC estimates for the decile score obtained in PP's main analysis sample (Sample B). PP's c-index estimates for the GRRS risk levels are 0.63 for whites; 0.62 for blacks; and 0.64 in the full sample. In contrast, as shown in Table 3.4, the AUC estimates for the more precise decile score in *Sample B* are 0.69 for whites; 0.69 for blacks; and 0.70 in the full sample.

AUC Results in Sample A

Table 3.3 shows the AUC results for the GRRS and VRRS in *Sample A* that includes defendants with complete records who have at least two years of follow-up time and that PP used to fit a reverse logistic regression predicting the "Not Low" Risk Level. For the GRRS analysis, *Sample A* is a subset of PP's two-year general recidivism data frame (*compas-scores-two-years.csv*) with filters applied as defined in the PP supplemental materials (*Compas Analysis.ipynb*).

For the GRRS decile scores, the AUC for whites is 0.693 $(0.670, 0.716)$. The AUC for blacks is 0.704 $(0.686, 0.722)$. The result of the test comparing the areas under the respective ROC curves for blacks and whites indicates the areas are not significantly different (p=0.438).

For the VRRS analysis, *Sample A* is a subset of the two-year violent recidivism data frame (*compas-scores-two-years-violent.csv*) with filters applied as defined in the PP supplemental materials. For the VRRS decile scores, the AUC for whites is 0.683 $(0.640, 0.726)$. The AUC for blacks is 0.708 $(0.680, 0.737)$. The result of the test comparing the areas under the respective ROC curves indicates the areas are not significantly different (p=0.383).

Table 3.3: AUC results for the General Recidivism Risk Scale (GRRS) decile scores and Violent Recidivism Risk Scale (VRRS) decile scores in the data analysis samples used for ProPublica's reverse logistic regression models (Sample A).

				base		Lower	Upper
	Sample	n	events	rate	AUC	95% CI	95% CI
GRRS							
	White	2103	822	0.39	0.693	0.670	0.716
	Black	3175	1661	0.52	0.704	0.686	0.722
	All	6172	2809	0.46	0.710	0.697	0.723
VRRS							
	White	1459	174	0.12	0.683	0.640	0.726
	Black	1918	404	0.21	0.708	0.680	0.737
	ALL	4020	652	0.16	0.719	0.698	0.741

Note. GRRS outcome is a misdemeanor or felony arrest. VRRS outcome is a violent misdemeanor or felony arrest.

AUC Results in Sample B

Table 3.4 shows the AUC results for the GRRS and VRRS in *Sample B* that consists of defendants with at least two years of follow-up and possibly incomplete case records. *Sample B* is used by PP to calculate the contingency tables from which they derive the main study results for false positive rates (fpr) and false negative rates (fnr). For the GRRS analysis, *Sample B* is a subset of PP's two-year general recidivism data frame (*compas-scores-two-years.csv*) with filters applied as defined in the PP supplemental materials (*Compas Analysis.ipynb*). For the GRRS decile scores, the AUC for whites is 0.693 $(0.672, 0.714)$. The AUC for blacks is 0.692 $(0.675, 0.709)$. The result of the test comparing the areas under the respective ROC curves indicates the areas are not significantly different (p=0.924).

Angwin et al. did not provide the data frame that they used to calculate the VRRS classification contingency tables. We constructed the data frame from the survival data provided on their web site (*cox-violent-parsed.csv*) with filters applied as defined in the PP supplemental materials, but our sample sizes did not perfectly match. Compared with PP's sample size reported in the VRRS classification tables, our data frame has 4 more black defendants (3,182 vs. 3,178); 5 fewer white defendants (2,260 vs. 2,265); and 1 less white failure (205 vs. 206). Some of the difference was due to cases with overlapping time intervals in their survival data (cox-violent-parsed.csv) that we excluded. For the VRRS decile scores, the AUC for whites is 0.656 $(0.616, 0.695)$. The AUC for blacks is 0.665 $(0.638, 0.692)$. The result of the test comparing the areas under the respective ROC curves indicates the areas are not significantly different (p=0.698).

Table 3.4: AUC results for the General Recidivism Risk Scale (GRRS) deciles scores and Violent Recidivism Risk Scale (VRRS) decile scores in the data analysis samples used for ProPublica's classification contingency tables (Sample B).

				base		Lower	Upper
	Sample	n	events	rate	AUC	95% CI	95% CI
GRRS							
	White	2454	966	0.39	0.693	0.672	0.714
	Black	3696	1901	0.51	0.692	0.675	0.709
	All	7214	3251	0.45	0.702	0.690	0.714
VRRS							
	White	2260	205	0.09	0.656	0.616	0.695
	Black	3182	443	0.14	0.665	0.638	0.692
	ALL	6454	735	0.11	0.681	0.660	0.701

Note. Compared with ProPublica's VRRS sample, our data frame has 4 more black defendants (3,182 vs. 3,178); 5 fewer white defendants (2,260 vs. 2,265); and 1 less white failure (205 vs. 206).

c-index Results in Sample C

Table 3.5 shows the c-index results for the GRRS and VRRS in *Sample C* that consists of 10,994 defendants with varying follow-up times (1 to 1,186 days) and was used by PP to fit the two Cox survival models. For the GRRS analysis, *Sample C* is a subset of PP's general recidivism survival data frame (*cox-parsed.csv*) with filters applied as defined in the PP supplemental materials (*Compas Analysis.ipynb*). For the GRRS decile scores, the *c*-index for whites is 0.658 $(0.641, 0.676)$. The *c*-index for blacks is 0.644 $(0.631, 0.657)$. The result of a test that the difference is greater than zero was not significant $(Pr(T \geq t) = .091)$.

For the VRRS analysis, *Sample C* is a subset of PP's violent recidivism survival data frame (*cox-violent-parsed.csv*) with filters applied as defined in the PP supplemental materials. For the VRRS decile score, the *c*-index for whites is 0.648 $(0.611, 0.685)$. The *c*-index for blacks is 0.651 $(0.625, 0.677)$. The result of a test that the difference is greater than zero was not significant $(Pr(T \geq t) = .560)$.

Table 3.5: C-Index results for the General Recidivism Risk Scale (GRRS) deciles scores and Violent Recidivism Risk Scale (VRRS) decile scores in the data analysis samples used for ProPublica's Cox Models (Sample C).

					Lower	Upper
	Sample	n	events	C-Index	95% CI	95% CI
GRRS						
	White	3569	1023	0.658	0.641	0.676
	Black	5147	2035	0.644	0.631	0.657
	All	10314	3469	0.664	0.654	0.674
VRRS						
	White	3826	221	0.648	0.611	0.685
	Black	5443	495	0.651	0.625	0.677
	ALL	10994	818	0.670	0.651	0.690

3.4 Checking the Results of PP's Survival and Logistic Regression Models

In this section we address the results from survival models that examined the interaction of race and risk level on the hazard of recidivism and PP's results from reverse logistic regression models that assessed the effect of race on risk level ("Low" vs. "Not Low").

Angwin et al. fit a Cox survival model to test the effects of the General Recidivism Risk Levels (Low, Medium, and High), race, and interactions of race by Risk Levels on the hazard of any arrest. For this analysis, there was a positive main effect for Black (coefficient = 0.2789) and large main effects for the Medium (0.843) and High (1.284) levels. The effect of Black was moderated by marginally significant interaction terms (Black:Medium = -0.173 and Black:High = -0.190). The predicted hazard ratios for the White group were 1.000, 2.323, and 3.609 for the Low, Medium, and High levels, respectively. The predicted hazard ratios for the Black group were 1.322, 2.583 and 3.945 for the Low, Medium, and High levels, respectively. Note that for all levels, black defendants are predicted to have a higher risk of recidivism than the white defendants. These results primarily show that there is a strong effect of COMPAS risk level for both the white and black groups and that the COMPAS levels underpredict for the black defendants. For example, based on this survival analysis, a black defendant who scores High will have a slightly higher predicted probability of recidivism than a white defendant who scores High. A reverse logistic regression analysis was carried out by the PP authors to show that the COMPAS levels overpredict for the Black defendants, which contradicts the results of this survival analysis.

The PP authors are essentially arguing that the COMPAS risk levels overpredict for blacks. If that were the case, then the predictions from a regression model that included the risk levels and race (black vs. white) designed to predict recidivism would need to be adjusted down for blacks vs whites. However, when Black is added to the Cox survival model, the predictions from the model for all three risk levels are adjusted in the opposite direction. Thus, the risk levels don't overpredict for blacks.

Angwin et al. also fit a Cox model to test the effects of the Violent Recidivism Risk Levels (Low, Medium, and High), race, and interactions of race by Risk Levels on the hazard of a violent arrest. They reported that the interaction terms in that model were not significant. We refit the model without the interaction terms and rechecked the main effect. The main effect for black was positive and nonsignificant. This finding indicates that the COMPAS levels do not underpredict or overpredict for blacks.

The PP authors fitted logistic regression models to test the effect of race on risk level ("Low", "Not Low"). This model supposedly controlled for prior crimes, future recidivism, age, and gender. Based on the results of these models, the PP authors claimed that "black defendants were 45% more likely to be assigned higher risk scores than white defendants" on the GRRS and "77% more likely to be assigned higher risk scores than white defendants" on the VRRS. These models are reverse logistic regressions. The standard practice for predictive models is to include future recidivism as the outcome in the model and the risk score as the predictor in the model. We argue that the reverse logistic regression models are misspecified because future recidivism is included as a predictor in the model and is measured with error. In addition, the results from these models are not consistent with the results from the Cox survival models. It is well-known that contradictory results can be obtained from reverse regression models (Greene, 1984). The standard way to test for race effects is to fit a model with recidivism as the outcome variable and risk score, race, and race by risk score as predictors, similar to the Cox survival models that the PP authors presented. The results from both of the Cox survival models (GRRS and VRRS) indicate that the risk levels do not overpredict for blacks, which contradicts the results in the reverse logistic regression models for the GRRS and VRRS.

Chapter 4

Conclusion

Our technical analysis was performed to determine the validity of the COMPAS Risk Scales for blacks and whites and to address the conclusions of Angwin et al. in the same Broward County data that they analyzed.

Angwin et al. give 5 major conclusions for their analysis. We enumerate each of these conclusions below and, for each conclusion, provide the correct conclusion, based on our technical review and analyses, supported in full by the data presented in this report. To arrive at the correct conclusions, we calculated the predictive values for blacks and whites in the target population.

1. PP Conclusion: "*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).*"

   Black defendants who were predicted to recidivate (i.e., given a "Not Low" score) actually did recidivate at a higher rate (63%) than the white defendants (59%). This finding provides evidence of predictive parity for the GRRS for blacks and whites in the target population.[1]

2. PP Conclusion: "*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).*"

   White defendants who were predicted not to recidivate (i.e., given a "Low" score) actually did not recidivate at a higher rate (71%) than the

[1] We use the Positive and Negative Predictive Values for stating the conclusions instead of their complements to follow standard practice in the field.

black defendants (65%). This finding provides evidence of predictive parity for the GRRS for blacks and whites in the target population.

3. PP Conclusion: "*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.*"

   The black defendants were not assigned inappropriately high risk scores. PP's conclusion is based on their results from their misspecified reverse logistic regression model, which is contradicted by their results from their correctly specified Cox survival model. PP's own Cox survival analysis showed that a variable coding White (0) versus Black (1) had a positive effect for predicting recidivism over and beyond the COMPAS Low, Medium, and High levels. There was an interaction showing that this effect was smaller for the Medium and High levels, but for all three levels, the predictions from the model were higher for black defendants. The PP authors are essentially arguing that the COMPAS risk levels overpredict for blacks. If that were the case, then the predictions from a regression model that included the risk levels and race (black vs. white) designed to predict recidivism would need to be adjusted down for blacks vs whites. However, when Black is added to the Cox survival model, the predictions from the model for all three risk levels are adjusted in the opposite direction. Thus, the risk levels don't overpredict for blacks. A more detailed technical discussion is provided in section 3.4.

4. PP Conclusion: "*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.*"

   Black defendants who were predicted to recidivate for violent recidivism (i.e., given a "Not Low" score) actually did recidivate at a marginally higher rate (21%) than the white defendants (17%). And white defendants who were predicted not to recidivate (i.e., given a "Low" score) actually did not recidivate at a marginally higher rate (93%) than the black defendants (91%). These findings provide evidence of predictive parity for the VRRS for blacks and whites in the target population.

5. PP Conclusion: "*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.*"

> The black defendants were not assigned inappropriately high risk scores for violent recidivism. In a survival analysis, a variable coding White (0) versus Black (1) had a positive but non-significant main effect for predicting recidivism over and beyond the COMPAS Low, Medium, and High levels. There was no interaction of this variable with levels as reported by PP. This finding indicates that the COMPAS risk levels neither underpredict or overpredict for black defendants.

In summary, Angwin et al. did not present any valid evidence that the risk scales are biased against blacks.

Appendix A

Decision Tables and ROC Curves

We demonstrate that the ROC curve is a device for examining performance across all thresholds of the risk scale. We show how the contingency tables for blacks and whites come from the ROC method and how the ROC curve is a plot of the tpr and fpr from contingency tables constructed at each threshold. We also show that the area under the ROC curve (AUC) is a summary of performance across all the thresholds of a risk scale. In this case the thresholds are the GRRS and VRRS decile scores. The analyses are carried out in study *Sample B* in which the PP authors calculate the false negative rate (fnr) and false positive rate (fpr) that they use to support the central claim of their article.

GRRS Classification Results by Race

Tables A.1 and A.2 show the *tradeoffs* across the thresholds of the GRRS for whites and blacks. At each threshold the tpr and fpr are calculated. This can be done by hand by constructing a contingency table like the one presented by Angwin et al. but at every decile score. The *Model Classification Statistics* are under the heading named "Model." The *Target Population Classification Statistics* are under the heading "Target Population." The study cut point ("Low" vs. "Not Low") is a decile score greater than or equal to 5.

It is a natural aspect of the fixed capacity of the receiver operating characteristic (ROC) method that moving the cut point of the classifier lower will increase the true positive rate (tpr) and increase the false positive rate (fpr) (Swets, 1988). Notice in Table A.1 if the cut point is moved from the Low threshold (decile score ≥ 5) to the High threshold (decile score ≥ 8), the tpr decreases from 0.52 to 0.20 and the fpr decreases from 0.23 to 0.05. There is a similar type of *tradeoff* for PV+ and PV-, as the location of the cut point is moved. As shown in Table A.1, if the cut point is moved from the Low threshold (≥ 5) to the High threshold (≥ 8), the positive predictive value (PV+) increases from 0.59 to 0.71, and the negative predictive value (PV-) decreases from 0.71 to 0.65.

Table A.1: Detail of Trade Offs Across the General Recidivism Risk decile scores for whites in the ProPublica study *Sample B* with two-year follow-up (n=2,454)

Deciles	**Model**		**Target Population**		
	true positive	false positive	PV+	PV-	Selection Ratio
(≥ 1)	1.00	1.00	0.39	1.00	1.00
(≥ 2)	0.85	0.64	0.46	0.79	0.72
(≥ 3)	0.74	0.47	0.50	0.76	0.58
(≥ 4)	0.64	0.35	0.54	0.74	0.46
(≥ 5)	0.52	0.23	0.59	0.71	0.35
(≥ 6)	0.41	0.15	0.64	0.69	0.25
(≥ 7)	0.29	0.09	0.68	0.66	0.17
(≥ 8)	0.20	0.05	0.71	0.65	0.11
(≥ 9)	0.12	0.03	0.70	0.63	0.07
(≥ 10)	0.05	0.01	0.70	0.61	0.03

Base Rate of Failure = 0.39

Angwin et al. limited their analysis to an examination of classification errors at the Low threshold (decile score ≥ 5). They misinterpreted these *Model Errors* (fnr, fpr) as if they were the *Target Population Errors* (1-NPV, 1-PPV). Here we present a proper interpretation of the classification statistics obtained at the study cut point.

As shown in Tables A.1 and A.2, there are four classification statistics to evaluate at this threshold: tpr (0.72) and fpr (0.45) for blacks and tpr (0.52) and fpr (0.23) for whites. The PP authors selectively reported and interpreted only the statistics that they thought supported their claim of racial bias against blacks. They failed to report that the comparison of the accuracy of the GRRS for blacks and whites at the study cut point ("Low", "Not Low") is inconclusive. Blacks have a higher fpr than whites (0.45 vs. 0.23), but blacks also have a higher tpr than whites (0.72 vs. 0.52).

We examine the difference more closely by comparing the relative classification statistics for blacks and whites at the study cut point.

Table A.2: Detail of Trade Offs Across the General Recidivism Risk decile scores for blacks in the ProPublica study *Sample B* with two-year follow-up (n=3,696)

Deciles	**Model**		**Target Population**		
	true positive	false positive	PV+	PV-	Selection Ratio
(≥ 1)	1.00	1.00	0.51	1.00	1.00
(≥ 2)	0.95	0.83	0.55	0.77	0.89
(≥ 3)	0.89	0.68	0.58	0.73	0.79
(≥ 4)	0.81	0.56	0.60	0.69	0.69
(≥ 5)	0.72	0.45	0.63	0.65	0.59
(≥ 6)	0.63	0.34	0.66	0.62	0.49
(≥ 7)	0.51	0.25	0.69	0.59	0.39
(≥ 8)	0.39	0.16	0.72	0.57	0.28
(≥ 9)	0.26	0.09	0.74	0.54	0.18
(≥ 10)	0.12	0.03	0.79	0.51	0.08

Base Rate of Failure = 0.51

The relative true positive rate (*rtpr*) of blacks vs. whites is $fpr_{white}/fpr_{black} = 0.523/0.720 = 0.726\ (0.679, 0.776)$. The relative false positive rate (*rfpr*) of blacks vs. whites at the study cut point is $tpr_{white}/tpr_{black} = 0.235/0.448 = 0.523\ (0.471, 0.581)$. If the (*rtpr*) is greater than 1 and the (*rfpr*) is less than 1, we would conclude that the accuracy of the classifier is higher for whites. The *rfpr* is less than 1, but the *rtpr* is not greater than 1. This is an inconclusive result. When the results of a comparison of the accuracy of a binary test in two groups is inconclusive, the results can be consolidated using a cost function (Pepe, 2003). That work is beyond the scope our report.

The relative negative predictive value (*rPV-*) of blacks vs. whites at the study cut point is $PV-_{white}/PV-_{black} = 0.712/0.650 = 1.09\ (1.04, 1.15)$. The relative positive predictive value (*rPV+*) of blacks vs. whites is $PV+_{white}/PV+_{black} = 0.591/0.630 = 0.939\ (0.880, 1.00)$. If both the *rPV-* and the *rPV+* are greater than 1, we would conclude that the risk scale performs better for whites. Our results show only that the *rPV-* is greater than 1. That is, the results indicate slightly better PV- for whites, but we cannot conclude that the classifier ("Low" vs. "Not Low") performs differently for blacks and whites.

Figure A.1 compares the smooth ROC curves of the General Recidivism Risk Scale decile scores for whites and blacks with at least two years of follow-up in the ProPublica study *Sample B*. The AUC is a summary of accuracy across all the thresholds of the GRRS decile score. Thus, the AUC summarizes the results of all the binary tests that could be conducted with the decile score. The AUC for whites is 0.693 (0.672, 0.714). The AUC for blacks is 0.692 (0.675, 0.709). The result of the test comparing the areas under the respective ROC curves indicates the areas are not significantly different (p=0.924). The *Model Errors* (tpr, fpr) at the study cut point for blacks and whites are printed on their respective ROC curves.

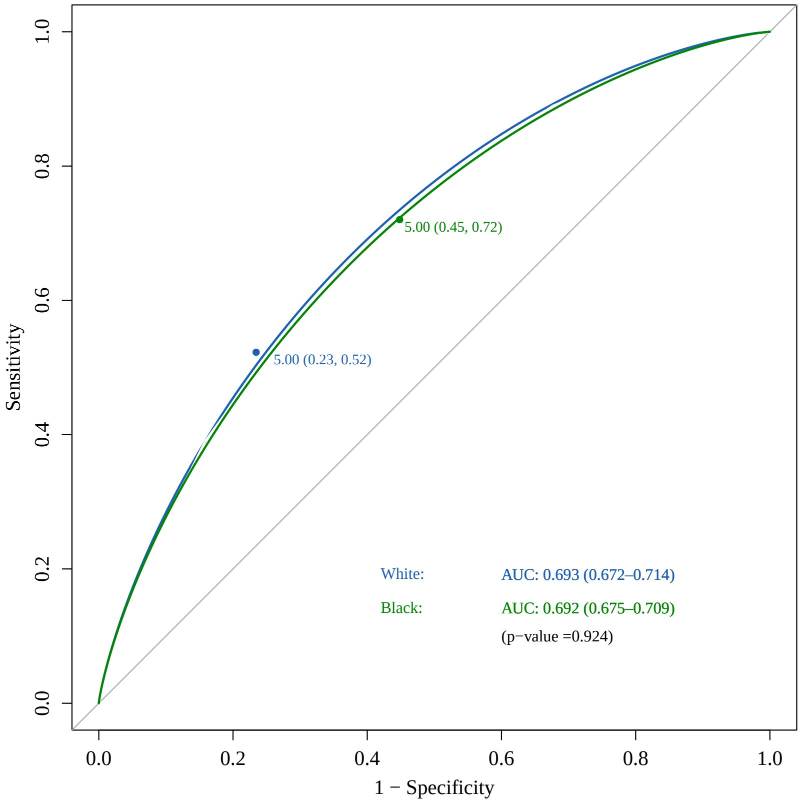


Figure A.1: Smooth ROC curves for the General Recidivism Risk Decile Score for whites (n=2,454) and blacks (n=3,696) in the ProPublica study *Sample B* with two-year follow-up.

VRRS Classification Results by Race

Tables A.3 and A.4 show the tradeoffs across the thresholds of the VRRS for whites and blacks. At each threshold the tpr and fpr are calculated. This can be done by hand by constructing a contingency table like the one presented by Angwin et al. but at every decile score. The *Model Classification Statistics* are under the heading named "Model." The *Target Population Classification Statistics* are under the heading "Target Population." The study cut point ("Low" vs. "Not Low") is a decile score greater than or equal to 5.

We illustrate the *tradeoffs* of the tpr and fpr using the results shown in Table A.3. If the cut point is moved from the Low threshold (decile score $\geq$ 5) to the High threshold (decile score $\geq$ 8), the tpr decreases from 0.38 to 0.11 and the fpr decreases from 0.18 to 0.03

Angwin et al. limited their analysis to an examination of classification errors at the Low threshold (decile score $\geq$ 5) of the VRRS. They compared the accuracy of the VRRS binary test ("Low" vs. "Not Low") for blacks and whites. There are four classification statistics to evaluate at this threshold: tpr (0.62) and fpr (0.38) for blacks and tpr (0.38) and fpr (0.18) for whites. The PP authors selectively reported and interpreted only the statistics that they thought supported their claim of racial bias against blacks. They failed to report that the VRRS classification statistics for blacks and whites were inconclusive. Blacks have a higher fpr than whites (0.38 vs. 0.18), but blacks also have a higher tpr than whites (0.62 vs. 0.38). We examine these differences more closely by comparing the relative accuracy and relative predictive values for blacks and whites.

The relative true positive rate (*rtpr*) of blacks vs. whites is $fpr_{white}/fpr_{black} = 0.376/0.616 = 0.610\ (0.503, 0.738)$. The relative false positive rate (*rfpr*) of blacks vs. whites is $tpr_{white}/tpr_{black} = 0.184/0.383 = 0.482\ (0.435, 0.534)$. If the (*rtpr*) is greater than 1 and the (*rfpr*) is less than 1, we would conclude that the accuracy of the classifier is higher for whites. The *rfpr* is less than 1, but the *rtpr* is not greater than 1. This is an inconclusive result.

There is a similar type of *tradeoff* for the PV+ and PV- as the location of the cut point is moved. As shown in Table A.3 if the cut point is moved from the Low threshold ($\geq$ 5) to the High threshold ($\geq$ 8), the positive predictive value (PV+) increases from 0.17 to 0.25 and the negative predictive value (PV-) decreases from 0.93 to 0.92. A comparison of the results in Table A.3 with the results in Table A.4, suggests slightly inconclusive *tradeoffs* for the predictive values for blacks and whites. The PV+ for blacks (0.21) is somewhat higher than the PV+ for whites (0.17), but blacks have a slightly lower PV- (0.91) in comparison with whites (0.93).

The relative negative predictive value (rPV-) of blacks vs. whites at the study cut point is $PV-_{white}/PV-_{black} = 0.169/0.207 = 0.817\ (0.65, 1.03)$. The relative positive predictive value (rPV+) of blacks vs. whites is $PV+_{white}/PV+_{black} = 0.929/0.909 = 1.02\ (1.00, 1.04)$. The confidence interval for the rPV- includes 1. Based on these results we cannot conclude that the classifier performs differently for blacks and whites.

Table A.3: Detail of Trade Offs Across the Violent Recidivism Risk decile scores for whites in the ProPublica study *Sample B* with two-year follow-up (n=2,260)

Deciles	Model		Target Population		
	true positive	false positive	PV+	PV-	Selection Ratio
(≥ 1)	1.00	1.00	0.09	1.00	1.00
(≥ 2)	0.78	0.57	0.12	0.95	0.59
(≥ 3)	0.65	0.41	0.14	0.94	0.43
(≥ 4)	0.51	0.27	0.16	0.94	0.30
(≥ 5)	0.38	0.18	0.17	0.93	0.20
(≥ 6)	0.29	0.12	0.20	0.93	0.13
(≥ 7)	0.19	0.06	0.23	0.92	0.07
(≥ 8)	0.11	0.03	0.25	0.92	0.04
(≥ 9)	0.07	0.02	0.28	0.91	0.02
(≥ 10)	0.01	0.01	0.14	0.91	0.01

Base Rate of Failure = 0.09

Table A.4: Detail of Trade Offs Across the Violent Recidivism Risk decile scores for blacks in the ProPublica study *Sample B* with two-year follow-up (n=3,182)

Deciles	Model		Target Population		
	true positive	false positive	PV+	PV-	Selection Ratio
(≥ 1)	1.00	1.00	0.14	1.00	1.00
(≥ 2)	0.93	0.80	0.16	0.94	0.82
(≥ 3)	0.85	0.65	0.17	0.94	0.68
(≥ 4)	0.72	0.51	0.19	0.92	0.54
(≥ 5)	0.62	0.38	0.21	0.91	0.42
(≥ 6)	0.52	0.27	0.23	0.90	0.31
(≥ 7)	0.38	0.18	0.25	0.89	0.21
(≥ 8)	0.26	0.10	0.30	0.88	0.12
(≥ 9)	0.17	0.06	0.32	0.88	0.07
(≥ 10)	0.06	0.02	0.35	0.87	0.02

Base Rate of Failure = 0.14

Figure A.2 compares the smooth ROC curves of the Violent Recidivism Risk Scale decile scores for whites and blacks with at least two years of follow-up in the ProPublica study *Sample B*. The AUC is a summary of accuracy across all the thresholds of the VRRS decile score. Thus, the AUC summarizes the results from all possible binary tests for the VRRS decile score. The AUC for whites is 0.656 (0.616, 0.695). The AUC for blacks is 0.665 (0.638, 0.692). The result of the test comparing the areas under the respective ROC curves indicates the areas are not significantly different (p=0.698). The *Model Errors* (tpr, fpr) at the study cut point for blacks and whites are printed on their respective ROC curves.

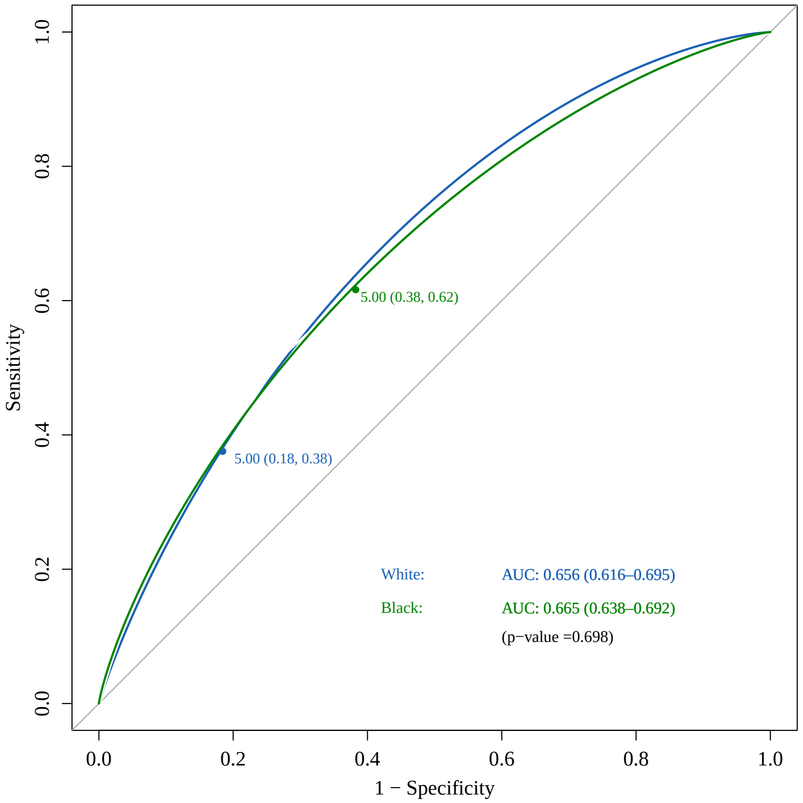


Figure A.2: Smooth ROC curves for the Violent Recidivism Risk Decile Score for whites (n=2,260) and blacks (n=3,182) in the ProPublica study *Sample B* with two-year follow-up.

Appendix B

Statistical and Other Technical Errors in the ProPublica Article

1. PP neglected to consider the base rate in the interpretation of their results. This is an error in judgement 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 (Kahneman & Tversky, 1982).

2. PP 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+).[3.2]

3. PP misrepresented the *Model Errors* as if they were *Target Population Errors*.

4. PP failed to report that the comparisons of the fpr and tpr for blacks and whites at the study cut point ("Low", "Not Low") for the VRRS and GRRS were inconclusive. Refer to Appendix A.

5. The reverse logistic regression models are misspecified. Refer to section 3.4.

6. The relative risk ratios from the reverse regressions are miscalculated and misinterpreted.

7. PP 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

8. PP misdefined the c-index as percent accuracy.

9. 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.

10. 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").

11. PP 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.

Appendix C

Glossary of Classification Terms

- *Performance = Diagnostic Accuracy (discriminative ability) + External Factors (base rate, user preferences, cut point, error cost ratios, selection ratio, treatment cost, treatment benefit)*

- *Receiver Operating Characteristic Method.* The ROC method is used to evaluate how accurately a risk scale discriminates between, for example, failures and successes or recidivists and non-recidivists. The term Receiver Operating Characteristic comes from engineering research where this type of analysis was first used to assess the accuracy of radar operators during World War II. Diagnostic characteristics are obtained by dividing a population into sub-samples of persons that recidivated (cases) and persons that did not recidivate (controls). At each cut point on the risk scale, a binary threshold or classifier is created. Those above the risk scale cut point are classified as recidivists, and we say they test positive. Those below the cut point are classified as non-recidivists, and we say they test negative.

  The test (risk scale classifier) is applied in the sample of recidivists (cases) and the percentage of cases that have a positive test result is calculated. This percentage is called the true positive rate. The true positive rate represents the *Sensitivity* of the test and is the probability of a positive test result in a sample of recidivists.

  Similarly the test (risk scale classifier) is applied in the sample of non-recidivists (controls), and the percentage of controls with a negative test result is calculated which is the true negative rate. The true negative rate represents the *Specificity* of the test and is the probability of a negative test result in a sample of non-recidivists.

  The Receiver Operating Characteristic (ROC) curve is a plot of Sensitivity and Specificity for all possible cut points. The area under the curve is a summary measure of accuracy across all the thresholds of the risk scale that was tested. Sensitivity and Specificity quantify *Model Accuracy*.

The complement of Sensitivity is the false negative rate which is the percentage of recidivists (cases) that have a negative test result. The complement of Specificity is the false positive rate which is the percentage of controls that have a positive test result. The false positive rate and false negative rate quantify *Model Error*.

Obviously these model statistics (operating characteristics) do not depend on the base rate of recidivism, because they are calculated separately on recidivists and non-recidivists. Note however that the trade-offs between Sensitivity and Specificity can be impacted by the base rate (see Bentley, Catanzaro, & Ganiaats, 2012, for example). Bentley et al. cite simulation and meta-analysis studies showing that an increase in prevalence will decrease the ratio of false to true positives and increase the positive predictive value, and that Sensitivity and Specificity can change as the base rate changes (see Chu, Nie, Cole, & Poole, 2009; Leeflang, Bossuyt, & Irwig, 2009, for example). Leeflang et al. (2013), using data from 23 meta-analyses to assess the effect of prevalence on Sensitivity and Specificity, found that Specificity decreased as disease base rate (prevalence) increased. In other words the false positive rate increased with increasing base rate.

- *Area Under the ROC Curve.* The AUC is a measure of the intrinsic accuracy of the risk scale (Swets, 1988). The examination of decision criteria and other factors external to the accuracy of the risk scales (prevalence, costs, benefits) provide information regarding the usefulness of the risk scales (Zweig & Campbell, 1993). The accuracy of a risk scale is defined by the AUC and is independent of decision criteria. The performance of a risk scale depends on the accuracy of the risk scale and also on decision criteria and prevalence (Harvey, 1992). When we talk of accuracy what we really mean is discrimination—the ability of a risk scale to discriminate between recidivists and non-recidivists. The AUC is a rank-based measure, not a measure of accuracy. Technically, accuracy refers to measures that take into account both calibration and discrimination. An example of a measure of accuracy is the Brier score. Note that the intrinsic accuracy of the risk scale decile score is not the same as the intrinsic accuracy of the raw risk scale score. Transformation to a decile score has in a sense introduced external criteria (norm-referenced cuts) that alter the shape of the ROC curve. The AUC for the decile score will generally be lower than the AUC for the raw score. Thus, the AUC calculated for the three-level risk score (Low, Medium, High) would be lower still.

The AUC is unrelated, or insensitive, to prevalence because both proportions that enter into its calculation are insensitive to prevalence. Other, perhaps more intuitive, measures of accuracy, such as overall proportion correct, are sensitive to prevalence and are therefore considered less useful measures of discrimination. For example, if the prevalence is only 5% one would expect to be correct 95% of the time with the no-fail classifier. The AUC value, on the other hand, can take on a range of values between 0.5 and 1 regardless of the prevalence and depends only on the intrinsic discriminating ability of the classifier.[1] An AUC value of 0.5 indicates classification performance no better than chance and a value of 1.0 indicates perfect classification. In the field of criminology, AUC values of 0.70 and higher are considered to be good.

- *Concordance Index.* The *c*-index is a generalization of the AUC. For binary data such as in logistic regression the *c*-index is identical to the AUC. The *c*-index is a measure of discrimination used to evaluate risk scales in survival data. The *c*-index is interpreted as the probability that the risk scores and survival times for a pair of randomly selected cases are concordant. A pair is concordant if the case with the higher risk score has a shorter survival time. The calculation is based on the number of all possible pairs of non-missing observations for which survival time can be ordered and the proportion of estimable pairs for which the predictors and survival times are concordant (Harrell, Califf, Pryor, Lee, & Rosati, 1982).

- *Predictive Value.* The positive predictive value (PV+) is the probability that a person with a positive test result will recidivate. The negative predictive value (PV-) is the probability that a person with a negative test result will not recidivate. Sensitivity and Specificity quantify the accuracy of the risk scale and the predictive value quantifies its clinical value (Pepe, 2003). A useful prediction will have a PV+ that is greater than the base rate and a PV- that is greater than 1 minus the base rate. A perfect test will predict the outcome perfectly with PV+ = 1 and PV- = 1. The predictive values depend on the accuracy of the test and the base rate of failure.

  The complement of the PV+ (1 - PV+) is the probability that a person with a positive test result will not recidivate. The complement of the PV- (1 - PV-) is the probability that a person with a negative test result will recidivate. These are the *Target Population Errors.* Linn (2004) discusses the difference between *Model Errors* (fpr and fnr from

[1] If values for AUC between 0 and 0.5 are observed, the predictions should be reversed, so that predictions of failures should be changed to predictions of not-to-fail.

the ROC method) and *Target Population Errors* (complements of predictive values). Referring to the *Model Errors*, Linn (2004) noted that these "statistics do not assess the accuracy of the diagnostic test in a clinically useful way, i.e., in a way that can be used by patients and physicians. Clinicians and patients relate to the sequelae of the clinical test results and do not know, at the time of performing the test, who has the disease and who does not (otherwise the test would not be performed). What is of interest to both physicians and patients is how many of the positively diagnosed patients, in fact, do not have the disease and how often a person with the disease is not diagnosed by the test."

- *Parity and Equity.* A risk scale exhibits *accuracy equity* if it can discriminate recidivists and non-recidivists equally well for two different groups such as blacks and whites. The risk scale exhibits *predictive parity* if the classifier obtains similar predictive values for two different groups such as blacks and whites, for example, the probability of recidivating, given a high risk score, is similar for blacks and whites. The interpretation of relative predictive values is discussed in Appendix A.

- *False Positive Paradox.* This is a classification result obtained for a test applied to a low base rate outcome where the probability of a False Positive (1-PV+) is high even though the test is accurate. The result goes against our intuition that tells us the probability of a False Positive (1-PV+) should be lower when the base rate of recidivism is lower. But in practice, for a given risk scale, we find that the lower the base rate of recidivism in the population, the more likely it is that an offender predicted to recidivate will not recidivate.

- *Base Rate Fallacy.* This is an error in judgement 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 (Kahneman & Tversky, 1982).

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). *Machine bias.* ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bentley, G. T., Catanzaro, A., & Ganiaats, T. G. (2012). Implications of the impact of prevalence on test thresholds and outcomes: Lessons from tuberculosis. *Biomed Central Research Notes*, *5*(563).

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, *36*, 21-40.

Chu, H., Nie, L., Cole, S. R., & Poole, C. (2009). Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Statistics in Medicine*, *28*(18), 2384–2399. Retrieved from http://dx.doi.org/10.1002/sim.3627 doi: 10.1002/sim.3627

Greene, W. H. (1984). Reverse regression: The algebra of disrimination. *Journal of Business & Economic Statistics*, *2*, 117-120.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association*, *247*, 2543-2546.

Harvey, L. O. (1992). The critical operating characteristic and the evaluation of expert judgment. *Organizational behavior and human decision processes*, *53*, 229-251.

Kahneman, D., & Tversky, A. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* (p. 153-160). New York, NY: Cambridge University Press.

Leeflang, M. M., Bossuyt, P. M., & Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: Implications for evidence-based diagnosis. *Journal of Clinical Epidemiology*, *62*, 5 - 12.

Leeflang, M. M., Rutjes, A. W., Reitsma, J. B., Hooft, L., & Bossuyt, P. M. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal*, *185*(11), E537-E544.

Linn, S. (2004). A new conceptual approach to teaching the interpretation of clinical tests. *Journal of Statistics Education*, *12*, 1 - 9.

Northpointe Inc. (2015a). Measurement and treatment implications of COMPAS Core Scales [Computer software manual]. Traverse City, MI.

Northpointe Inc. (2015b). *Practitioner's guide to COMPAS Core.* Traverse City, MI. Retrieved from http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford University Press.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285-1293.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561-577.