Harvard Data Science Review • Issue 2.1, Winter 2020

The Age of Secrecy and Unfairness in Recidivism Prediction

Cynthia Rudin, Caroline Wang, Beau Coker

Published on: Mar 31, 2020

DOI: 10.1162/99608f92.6ed64b30

ABSTRACT

In our current society, secret algorithms make important decisions about individuals. There has been substantial discussion about whether these algorithms are unfair to groups of individuals. While noble, this pursuit is complex and ultimately stagnating because there is no clear definition of fairness and competing definitions are largely incompatible. We argue that the focus on the question of fairness is misplaced, as these algorithms fail to meet a more important and yet readily obtainable goal: transparency. As a result, creators of secret algorithms can provide incomplete or misleading descriptions about how their models work, and various other kinds of errors can easily go unnoticed. By trying to partially reconstruct the COMPAS model—a recidivism risk-scoring model used throughout the criminal justice system—we show that it does not seem to depend linearly on the defendant's age, despite statements to the contrary by the model's creator. This observation has not been made before despite many recently published papers on COMPAS. Furthermore, by subtracting from COMPAS its (hypothesized) nonlinear age component, we show that COMPAS does not necessarily depend on race other than through age and criminal history. This contradicts ProPublica's analysis, which made assumptions about age that disagree with what we observe in the data. In other words, faulty assumptions about a proprietary model led to faulty conclusions that went unchecked until now. Were the model transparent in the first place, this likely would not have occurred. We demonstrate other issues with definitions of fairness and lack of transparency in the context of COMPAS, including that a simple model based entirely on a defendant's age is as 'unfair' as COMPAS by ProPublica's chosen definition. We find that there are many defendants with low risk scores but long criminal histories, suggesting that data inconsistencies occur frequently in criminal justice databases. We argue that transparency satisfies a different notion of procedural fairness by providing both the defendants and the public with the opportunity to scrutinize the methodology and calculations behind risk scores for recidivism.

Keywords: transparency in predictive modeling, criminal justice, risk assessment, machine learning, trustworthiness

Media Summary

In our current society, increasingly we rely on secret algorithms to make important decisions about individuals. In criminal justice, there has been substantial concern about due process, and whether these secret algorithms are unfair. One such secret algorithm—called COMPAS—is widely-used across the justice system, and has been the subject of high-profile lawsuits. But what do secret algorithms like COMPAS actually compute? If we could peek inside, perhaps we could determine what makes these algorithms tick—and whether we need them at all.

This article reports our effort to use publicly available data from the justice system to infer parts of the COMPAS formula. Our analysis indicates that the secret formula is not what some people thought. In particular, what we uncovered runs counter to some claims made by ProPublica about what is inside this secret algorithm. Our analysis also suggests that COMPAS scores may often be miscomputed. These kinds of errors can lead to years of extra prison time, or the other extreme, dangerous individuals being released into society. These findings draw attention to how important transparency is in judicial decision making when using AI, machine learning, algorithms, or other types of statistical models. More specifically, these results underscore the dangers of judicial decision-making based on secret risk-assessment algorithms, as opposed to transparent alternatives.

This article is accompanied by multiple invited discussion pieces and a rejoinder by the author.

---

1. Introduction

Secret algorithms control important decisions about individuals, such as judicial bail, parole, sentencing, lending decisions, credit scoring, marketing, and access to social services. These algorithms may not do what we think they do, and they may not do what we want.

There have been numerous debates about fairness in the literature, mainly stemming from a flawed analysis by the ProPublica group (Larson, Mattu, Kirchner, & Angwin, 2016; Angwin, Larson, Mattu, & Kirchner, 2016) of data from Broward County, Florida, claiming that the proprietary prediction model COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) (Brennan, Dieterich, & Ehret, 2009) is racially biased. COMPAS is used throughout the criminal justice system in the U.S., and its predictions have serious consequences in the lives of many people (Baradaran, 2013; Citron, 2016; Corbett-Davies, Pierson, Feller, & Goel, 2016; Crow, 2008; Flores, Lowenkamp, & Bechtel, 2016; Gottfredson & Jarjoura, 1996; Lowenkamp & Latessa, 2004; Netter, 2007; Pennsylvania Commission on Sentenceing, 2018; Petersilia & Turner, 1987; Redding, 2009). The bottom line from these debates is that there is not a single correct definition of fairness and that multiple types of fairness are incompatible. We put aside typical fairness considerations for a moment to focus on more pressing issues.

One issue with COMPAS is that it is *complicated*. It is based on up to 137 variables (Northpointe, 2009) that are collected from a questionnaire. This is a serious problem because typographical or data entry errors, data integration errors, missing data, and other types of errors abound when relying on manually entered data. Individuals with long criminal histories are sometimes given low COMPAS scores (which labels them as low risk), and vice versa. In the past, there have been documented cases where individuals have received incorrect COMPAS scores based on incorrect criminal history data

(Wexler, 2017a, 2017b) and have possessed no mechanism to correct it after a decision was made based on that incorrect score. This problem has inadvertently occurred with other (even transparent) scoring systems, in at least one case leading to the release of a dangerous individual who committed a murder while on bail (Ho, 2017; Westervelt, 2017). An error in a complicated model is much harder to find than an error in a simple model, and it not clear how many times typographical errors in complicated models have led to inappropriate releases that resulted in crimes, after decades of widespread use of these models. The question of whether calculation errors occur often in these models is of central importance to the present work.

A separate issue with COMPAS is that it is *proprietary*, which means its calculations cannot be double-checked for individual cases, and its methodology cannot be verified. Furthermore, it is unclear how the data COMPAS collects contribute to its automated assessments. For instance, while some of the questions on the COMPAS questionnaire are the same as those in almost every risk score—age and number of past crimes committed—other questions seem to be direct proxies for socioeconomic status, such as "How hard is it for you to find a job ABOVE minimum wage compared to others?"[1] It is not clear that such data should be collected for the purposes in which these risk scores are used.

Though creators of proprietary algorithms often provide descriptions of how their models work, by nature, it is difficult for third parties to verify these descriptions. This may allow errors in documentation to remain undiscovered for years. By partially reconstructing COMPAS in Broward County, we show in Section 2.2 that COMPAS may depend nonlinearly on age, contradicting its stated methodology.

ProPublica, who did not know how COMPAS depends on age, assumed a specific form for age, and concluded that being African American leads to a higher COMPAS score, even controlling for criminal history and sex. Their assumption appears to be incorrect, and thus their conclusions were invalid. While adding the correct nonlinear age term to the regression could mitigate the impact of this particular issue, it misses the larger point. Without transparency, an incorrect understanding of a model (e.g., the form of its dependence on age) can go unchecked, leading to downstream consequences for independent analyses of a model.

While COMPAS depends heavily on age, we show in Sections 2.3 through 2.6 that it does not seem to depend strongly on either criminal history or proxies for race. That is, it is possible that COMPAS depends less on criminal history than we might expect. This leads to the possibility that COMPAS instead depends heavily on variables that we may not want it to depend on.

In Section 3 we pinpoint many individuals whose COMPAS scores seem unusually low given their criminal histories. Since COMPAS is proprietary, we cannot fully determine whether these low scores

are due to errors in calculation, data entry errors, or errors from some other source (or even if they are errors at all).

COMPAS's creator Northpointe disagreed with each of ProPublica's claims on racial bias based on their definition of fairness (Dieterich, Mendoza, & Brennan, 2016). Their rebuttal did not include arguments as to the type of fairness we consider here, and in particular, why it benefits the justice system to use a model that is complicated or proprietary.

Work in machine learning has shown that complicated, black-box, proprietary models are not necessary for recidivism risk assessment. Researchers have shown (on several datasets, including the data from Broward County) that interpretable models are just as accurate as black box machine learning models for predicting recidivism (Zeng, Ustun, & Rudin, 2017; Tollenaar & van der Heijden, 2013; Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2017; Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2018; Rudin & Ustun, 2018; Ustun & Rudin, 2019). These simple models involve age and counts of past crimes, and indicate that those who are younger or have longer criminal histories are more likely to reoffend. A judge could easily memorize the models within these works, and compute the risk assessments without even a calculator (Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2017; Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2018). Despite this knowledge, proprietary models are still being used.

Given that we do not need proprietary models, why we should allow proprietary models at all? The answer is the same as it is in any other application: by protecting intellectual property, we incentivize companies to perform research and development. Since COMPAS has been at the forefront of the fairness debate about modern machine learning methods, it is easy to forget that COMPAS is not one of these machine learning methods. It is a product of years of painstaking theoretical and empirical sociological study. For a company like Northpointe to invest the time and effort into creating such a model, it seems reasonable to afford the company intellectual property protections. However, as we discussed, machine learning methods—either standard black box or, better yet, recently developed interpretable ones—can predict equally well or better than bespoke models like COMPAS (Zeng, Ustun, & Rudin, 2017; Tollenaar & van der Heijden, 2013; Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2017; Angelino, Larus-Stone, Alabi, Seltzer, & Rudin, 2018). For important applications like pre-trial release or parole decisions in criminal justice, academics have always been willing to devote their time and energy. High-performing predictive models can therefore be created with no cost to the criminal justice system. Allowing proprietary models to incentivize model development was never necessary in the first place.

Neglecting to use transparent models has consequences. We provide two arguments for why transparency should be prioritized over other forms of fairness. First, no matter which technical definition of fairness one chooses, *it is easier to debate the fairness of a transparent model than a*

proprietary model. Transparent models provide defendants and the public with imperative information about tools used for safety and justice, allowing a wider audience to participate in the discussion of fairness. Second, *transparency constitutes its own type of procedural fairness* that should be seriously considered (Coglianese & Lehr, 2018). We argue that it is not fair that life-changing decisions are made with an error-prone system, without entitlement to a clear, verifiable, explanation.

In Section 2, we try to partially reconstruct COMPAS for Broward County and show how it is likely to be inconsistent with its official documentation; in Section 3, we identify a number of individuals with long criminal histories but low risk scores; and in Section 4, we describe transparency as a form of fairness. We consider the most transparent nontrivial predictive model we could find: age. Younger people tend to be at higher risk of recidivism. Our goal in this section is to modify the discussion of fairness to be through the lens of transparency.

2. Reconstructing COMPAS

Even with our limited data, we may have succeeded in partially reconstructing parts of the COMPAS model as it is implemented in Broward County. We will next describe these attempts. Note that other groups besides ProPublica have attempted to explain COMPAS (Tan, Caruana, Hooker, & Lou, 2018; Stevenson & Slobogin, 2018). One attempt, by Stevenson and Slobogin (2018) uses only linear models (and our evidence suggests that COMPAS's dependence on age is nonlinear). Another attempt, that of Tan et al. (2018), modeled COMPAS with a generalized additive model, and we hypothesize that they chose the correct model form, based on COMPAS's documentation. However, their analysis has similar issues to that of ProPublica or that of Stevenson and Slobogin (2018): all of these groups attempted to explain COMPAS scores using ProPublica's limited set of features, but this type of analysis is not valid because there are many unmeasured features. COMPAS's actual dependence on the observable features may be totally different than what they report. For instance, if age is correlated with unmeasured survey data, their model would exploit this correlation to compensate for the missing survey data, leading to an incorrect estimate of how COMPAS depends on age. Our approach instead attempts to isolate and subtract off the parts of COMPAS that we think can be explained from our data, using COMPAS' documentation to guide us. If our stated assumptions are correct, our analysis is valid even in the presence of many unmeasured features. The analysis still holds even if the measured and unmeasured features are correlated. Of course, ground truth cannot be made available (except to the designers of COMPAS), so our hypotheses cannot be verified.

2.1 COMPAS as Described by Its Creator

There are two COMPAS recidivism risk assessments of interest: the general score and the violent score. These scores assess the risk that a defendant will commit a general or violent crime, and we use them to predict crime within the next two years. Each score is given as an integer between 1 and 10 but

is based on a raw score that can take any value. Higher scores indicate higher risk. The raw score is computed by a formula and the final integer score is normalized based on a local population. We will therefore attempt to reconstruct the raw scores.

To compute the COMPAS raw scores, Northpointe collects 137 variables from a questionnaire, computes a variety of *subscales*, and finally linearly combines the subscales and two age variables — the defendant's age at the time of the current offense and age at the time of the first offense — to compute the raw risk scores. For example, using the equation exactly as written in the COMPAS documentation (Northpointe, 2019), the violent recidivism raw score is given by:

$$\begin{aligned}\text{Violent Recidivism Risk Score} \\ = \quad & (\text{age} * -\text{w}) + (\text{age-at-first-arrest} * -\text{w}) + (\text{history of violence} * \text{w}) \\ & + (\text{vocation education} * \text{w}) + (\text{history of noncompliance} * \text{w}),\end{aligned}$$

where the variables not related to age are subscales and the weights w may be different for each variable. The notation age $*-$w would commonly indicate "age multiplied by the negative of w." We have little knowledge of how the subscales depend on the questionnaire; the documentation states only which questionnaire items are used for which subscales. Table 1 shows for each subscale the recidivism score(s) to which it relates and the number of underlying questionnaire items we can compute using our data. We use the data made available by ProPublica;[2] the ProPublica data set is missing features needed to compute the COMPAS score, and so we supplement this data set with probation data from the Broward Clerk's Office. However, there remain missing items often related to subjective survey questions that cannot be computed without access to Northpointe's data, which are not publicly available. Notes on our data processing can be found in the Appendix.

Subscale	**# Features we have/Total**	**Relevant Recidivism Score**
Criminal Involvement	4/4	General
History of Noncompliance	3/5	Violent
History of Violence	8/9	Violent
Vocational/Educational	0/12	Both

Substance Abuse	0/10	General

Table 1. COMPAS Subscales Are Inputs to the Recidivism Scores.[3]

According to our understanding of the COMPAS documentation, the general and violent COMPAS scores are both linear models, where age and age-at-first-arrest are the only factors that can have negative contributions to the COMPAS scores.

2.2 COMPAS Seems to Depend Nonlinearly on Age, Contradicting its Documentation

Let us consider whether the linear model form given by COMPAS's documentation is supported by our data. We make the following assumption:

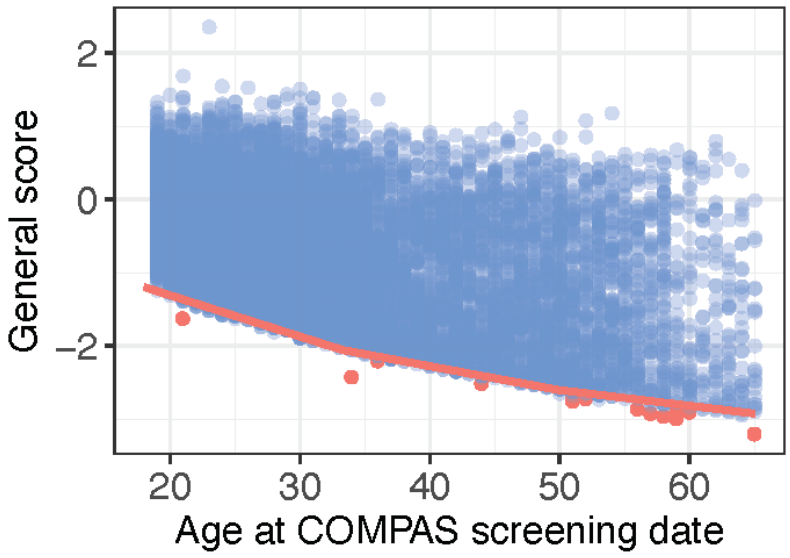
> **Data Assumption**: *For most values of age, there is at least one person in our dataset with age-at-first-arrest equal to their age and the lowest possible risk score for each subscale.*

First, note that age and age-at-first-arrest are the only COMPAS score inputs that have negative contribution to the COMPAS scores. Next, note that the least-risky value for age-at-first-arrest occurs when it is equal to current age, since this implies individuals committed their first arrestable offense at their current age. Thus, individuals satisfying the Data Assumption should have the lowest COMPAS raw score for their age, which is key for the rest of our age analysis. In what follows, we describe attempts to confirm that these individuals are present in the data.

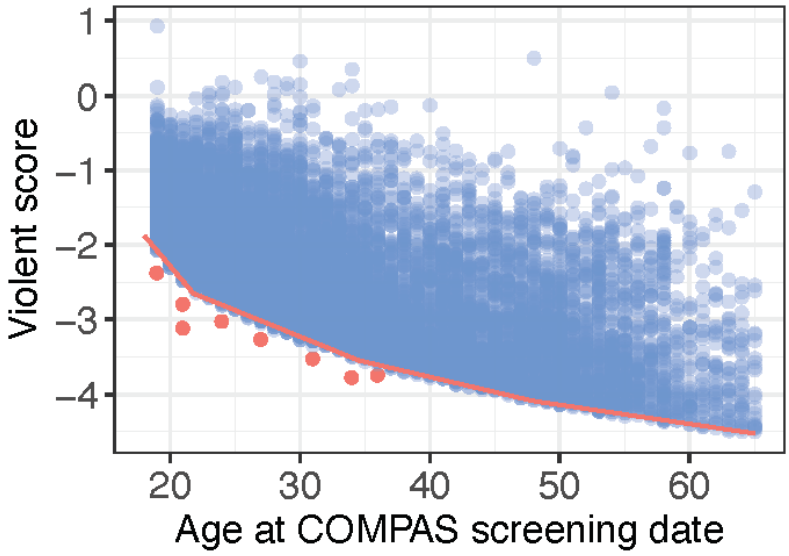
First, and most importantly, we used our data to check directly that for most values of age, there is at least one person in our dataset who has age-at-first-arrest equal to their age and who does not have criminal history. For the COMPAS general score, we can check if the Data Assumption holds for the Criminal Involvement Subscale because we have all the inputs to it. We can only approximately check this assumption for the History of Violence and History of Noncompliance Subscales because we do not possess all the inputs. However, the Data Assumption seems plausible given that (1) the inputs to the subscales can take only a few integer values[4] and (2) their typical content (e.g., 'Do you have a job?' or 'How many times has this person been sentenced to jail for 30 days or more?') suggests the least-risky input values are fairly likely. We cannot check the Data Assumption for the subscales for which we do not have data. However, the Data Assumption seems to hold for the subscales for which we *do* have data (see Figure A3 in the Appendix), leading us to believe it might hold generally.

If the COMPAS documentation were correct in that the COMPAS models are linear functions of age, then as long as the Data Assumption holds, if we plot the COMPAS raw scores against age for all

people, the lower bounds should be a line with slope equal to the sum of the coefficients on age and age-at-first-arrest, since age equals age-at-first-arrest. Figure 1 shows this is not the case. Also, the people near the lower bounds of Figure 1 often have no criminal history, and have age-at-first-arrest equal to age (see the Appendix for more analysis). Thus, our evidence indicates the COMPAS models are not the linear models given within the documentation. Recall that except for some typos, there should be no noise in the COMPAS scores. The COMPAS scores we observe should be the direct output of a computer program.



f_{age}



$f_{\text{viol age}}$

Figure 1. Scatterplot of COMPAS general recidivism score versus age and scatterplot of COMPAS violent recidivism score versus age. Age splines used to approximate the lower bound of the scatterplots are shown in red; age outliners that were removed from the analysis are also in red.

Despite the lack of agreement of the data with the documentation, we would still like to reconstruct as much of the COMPAS black box as possible, so we make several weaker hypotheses about its form, none of which contradict the COMPAS documentation:

> **Model Assumption 1**: *Both COMPAS raw scores are additive models with respect to each input (age, age-at-first-arrest, and the subscales).*
>
> **Model Assumption 2**: *The additive terms corresponding to each input except age and age-at-first-arrest are never negative.*
>
> **Model Assumption 3**: *Given age, the lowest risk score is attained when (a) the additive term corresponding to age-at-first-arrest is lowest (i.e., when age-at-first-arrest is as large as possible; that is, equal to age) and (b) the additive terms corresponding to the subscales are zero.*

We believe Model Assumption 2 should hold because all of the inputs except age and age-at-first-arrest (e.g., number of violent felonies, number of times on probation) should lead to a higher risk of violence or recidivism, and therefore a higher COMPAS score. Should Model Assumption 3 not hold, people with nonzero criminal history would have lower COMPAS scores than those with none, which is

not intuitive. With these Model Assumptions and under the Data Assumption, the lower bound observed in Figure 1 is exactly the additive term corresponding to age.

Reconstructing COMPAS's dependence on age is important because we know that COMPAS, if it is like other scoring systems, should depend heavily on age in order to predict well. If we could isolate and subtract off its dependence on age, we could more easily determine its dependence on protected attributes such as criminal history and race. Based on the lower bound of the COMPAS score with respect to age, we present a conjecture of approximately how the COMPAS score may depend on age, at least in Broward County, and we have a similar conjecture for the violent recidivism counterpart:

Conjecture: *The COMPAS general recidivism model is a nonlinear additive model. Its dependence on age in Broward County is approximately a linear spline, defined as follows:*

$$\begin{aligned} \text{for ages } \leq 33.27, &\quad f_{\text{age}}(age) = -0.056 \times age - 0.181 \\ \text{for ages between } 33.27 \text{ and } 49.82, &\quad f_{\text{age}}(age) = -0.033 \times age - 0.961 \\ \text{for ages } \geq 49.82, &\quad f_{\text{age}}(age) = -0.021 \times age - 1.534. \end{aligned}$$

Similarly, the COMPAS violent recidivism model is a nonlinear additive model, with a dependence on age that is approximately a linear spline, defined by:

$$\begin{aligned} \text{for ages } \leq 21.77, &\quad f_{\text{viol age}}(age) = -0.205 \times age + 1.815 \\ \text{for ages between } 21.77 \text{ and } 34.51, &\quad f_{\text{viol age}}(age) = -0.070 \times age - 1.114 \\ \text{for ages between } 34.51 \text{ and } 47.91, &\quad f_{\text{viol age}}(age) = -0.040 \times age - 2.151 \\ \text{for ages } \geq 47.91, &\quad f_{\text{viol age}}(age) = -0.026 \times age - 2.865. \end{aligned}$$

We used a two-stage procedure to fit the functions f_{age} and $f_{\text{viol age}}$, with the goal of obtaining the closest approximation to the true functional relationship between age and the COMPAS raw scores as possible. First, we fit quadratic and quartic polynomials respectively to the lower bounds of the scatterplots of individuals' COMPAS general recidivism scores and the COMPAS violent recidivism scores. Points more than $c = .05$ below these age polynomials were deemed "age outliers" (a handful of individuals whose age seems to be incorrect) and removed from the analysis. Fewer than 10 individuals (for each score) were removed due to this reason. The age splines were fit using the lower bound of the subsets of individuals whom we hypothesize to satisfy the data assumption (individuals with age equal to age-at-first-arrest and no known contribution to any subscale), shown in Figure 2. In the left figure, several of the age outliers appear as if they were in a line, but if we add in the age

outliers from the full population (not just those with no criminal history) the apparent line is no longer a lower bound.

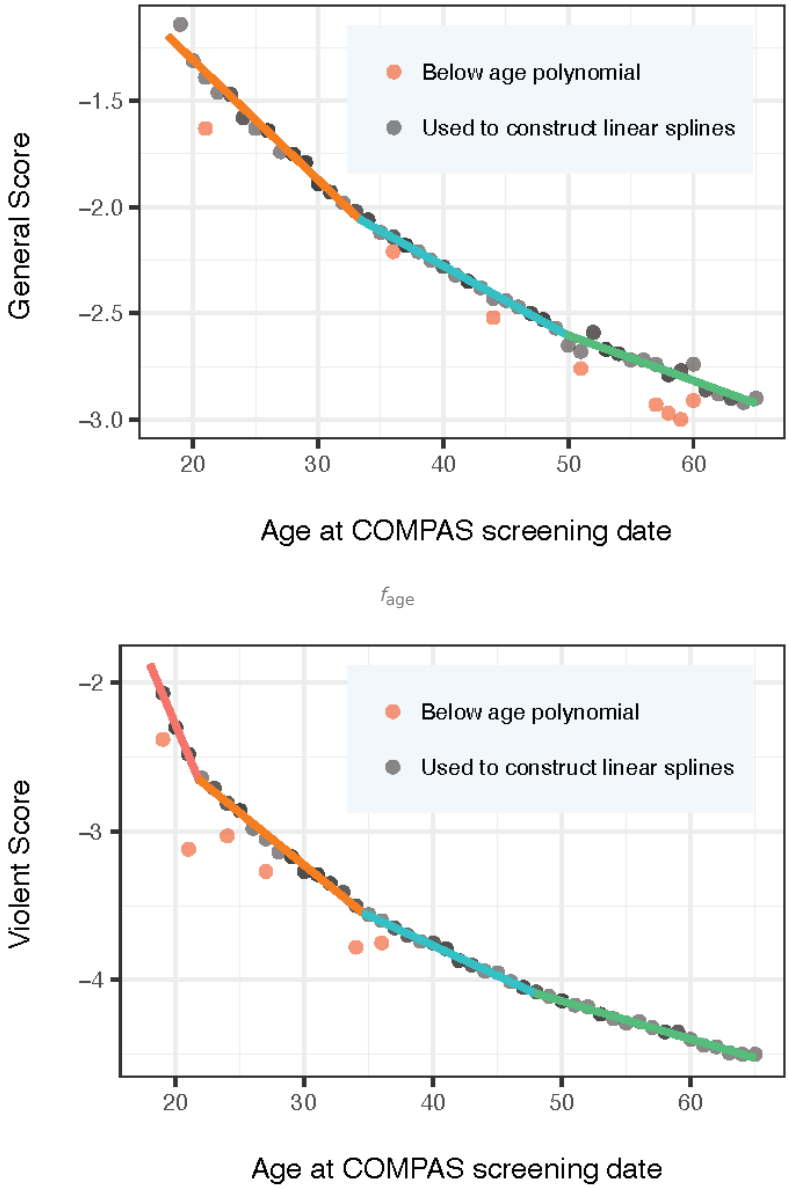


Figure 2. Fitting f_{age} and $f_{viol\ age}$ to the data. The data are represented by translucent gray dots (overlapping data points appear darker) and the fitting is shown as the multicolored line. Points deemed as age outliers are shown in red. Ninety-two individuals were used to fit $f_{viol\ age}$ and 100 individuals were used to fit f_{age}.

As discussed above, COMPAS's documentation claims a linear dependence on age. Even if COMPAS constructed the model to only take age as a linear argument, its predictions, as applied to a sample

with representative covariate information, apparently induce nonlinear dependence on it. For the COMPAS documentation to claim such linear dependence can thus be misleading.

It is possible that age's *only* contribution to the COMPAS general recidivism score is f_{age} (similarly, $f_{\text{viol sage}}$ for the violent score). Let us describe why this seems to be true. *The remainders of general COMPAS minus* f_{age} *and violent COMPAS minus* $f_{\text{viol age}}$ *do not seem to depend on age.* After subtracting out the age polynomials, we employed machine learning methods along with linear models (Table 2) to model the remainder (COMPAS score minus the age polynomial). We ran each of these algorithms on the data, once using criminal history and age features only (*with-age* models), and once using just criminal history (*without-age* models). Machine learning methods are powerful, nonparametric models that are capable of modeling nonlinear functions very closely, given the right features. Thus, if the remainder depends on age coupled with criminal history, it is likely the accuracy will vary between the with-age and without-age models. However, instead, Tables 2 and 3 (for the general and violent scores, respectively) show the accuracy of the machine learning models was almost identical between the with-age and without-age models.

Importantly, if the dependence on age is additive, COMPAS does not use features that couple age and criminal history, such as the rate of crimes committed over time, despite the potential usefulness of this type of feature (see, e.g., Bushway & Piehl, 2007).

	Linear Model	**Random Forest**	**Boosting**	**SVM**
Without Age	0.565	0.528	0.513	0.523
With Age	0.562	0.525	0.507	0.518

Table 2. Root Mean Square Error for Several Machine Learning Algorithms for Predicting COMPAS General Score Minus Age Polynomial (f_{age}), With Age Included as a Feature (bottom row), and Without Age (top row).[5]

	Linear Model	**Random Forest**	**Boosting**	**SVM**
Without Age	0.471	0.460	0.453	0.462
With Age	0.463	0.447	0.439	0.447

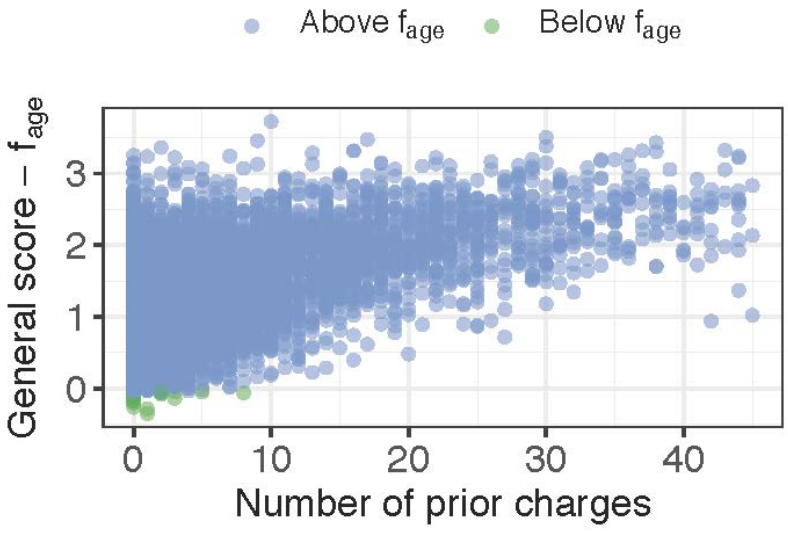
Table 3. Analogous to Table 2 but for the COMPAS Violent Score, Predicting the COMPAS Violent Score Minus $f_{\text{viol age}}$. [6]

The fact that the lower bounds f_{age} and $f_{\text{viol age}}$ seem to vary smoothly and uniformly with age, with only few outliers, indicates that the data entering into the COMPAS scores is high quality with respect to age. This has implications for our later analysis.

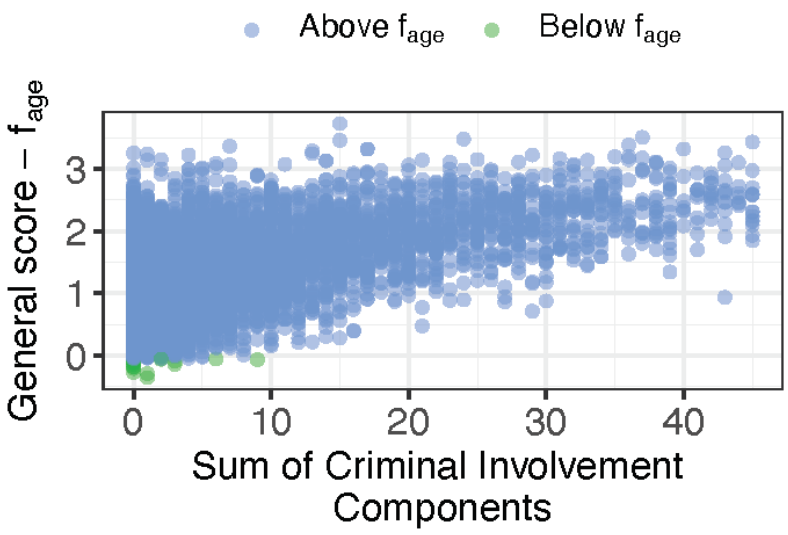
Now that we have hypothesized the dependence of COMPAS on age, we wish to explain its dependence on criminal history variables. We do this separately for the general and violent scores in Sections 2.3 and 2.4, respectively.

2.3 Criminal History and the COMPAS General Recidivism Score

Unlike the dependence on age, the COMPAS general score does not seem to display a clear dependence on criminal history. Figure 3 shows a scatterplot of COMPAS general score *remainder* (which we define as the COMPAS score minus the age spline f_{age}) against the total number of prior charges, which is one of the variables determining the Criminal Involvement Subscale (left panel), and the unweighted sum of the variables in the Criminal Involvement Subscale (right panel). Note that we would ideally plot the remainder against the Criminal Involvement Subscale itself, but we do not know how the inputs are combined to form the subscale. Even excluding the age outliers (highlighted in green), there is no smooth lower bound as seen in Figure 1. Therefore we transition from searching for simple dependence of the COMPAS general score on its subscale items to searching for more complex dependencies.



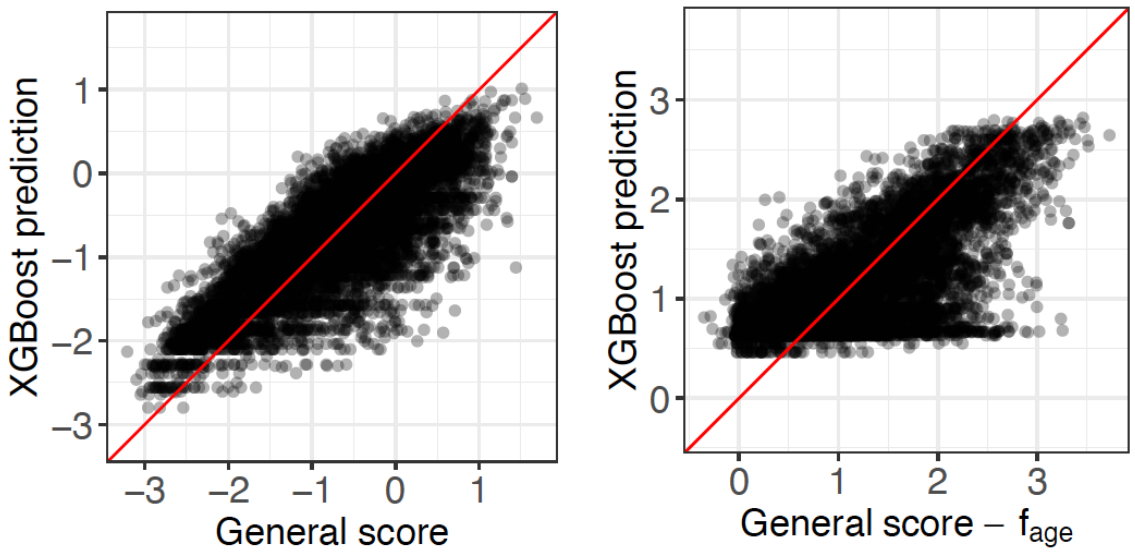
COMPAS $-f_{\text{age}}$ vs. number of prior charges. The green points are age outliers.



COMPAS $-f_{age}$ vs. unweighted sum of Criminal Involvement Subscale components. The green points are age outliers.

Figure 3. We do not find a clear relationship between the COMPAS general score after subtracting the age spline and criminal history information. Note that in each plot there are a few observations with a large number of prior charges that are outside of the plot range. Left: COMPAS− f_{age} vs. number of priors. The green points are age outliers. Right: COMPAS− f_{age} vs. unweighted sum of Criminal Involvement Subscale components

To then investigate whether the COMPAS general score depends in a more complex way on the Criminal Involvement Subscale items listed in the Appendix in Table A6, we ran several machine learning algorithms (random forests, Breiman, 2001, boosted decision trees, Freund & Schapire, 1997, and support vector machines with a radial basis kernel function, Vapnik, 1998) on the subscale items our data contains, to see if the COMPAS general recidivism score could be explained (either linearly or nonlinearly) by the subscale components. We tried predicting both the general score itself and the general score after subtracting f_{age}. Figure 4 shows a scatterplot of predictions versus the actual values for the two prediction problems. We make two observations from this figure. By comparing the two panels, we can see that the COMPAS general score seems to depend heavily on age, as the predictions of the COMPAS score remainder (right panel) are much worse than the predictions of the COMPAS score itself (left panel); this is because criminal history is correlated with age. After subtracting our reconstructed dependence on age (right panel), we see the ability of the criminal history variables to predict the COMPAS score remainder is surprisingly *un*successful. *Thus the dependence of the COMPAS general score on criminal history, as captured by the components of the Criminal Involvement Subscale, seems to be weak.*



(a) Predictions of COMPAS general score vs. actual values.

(b) Predictions of COMPAS general score $-f_{\text{age}}$ vs. actual values.

Figure 4. Predicting general remainder.

2.4 Criminal History and the COMPAS Violent Recidivism Score

We gained more traction reconstructing the COMPAS violent recidivism score than the general score. Figure 5 shows the COMPAS violent score after subtracting the age spline $f_{\text{viol age}}$ against the unweighted sum of the Violence History Subscale components. Excluding the age outliers, this subtraction produced a crisp lower bound on the remainder, unlike the bounds we obtained trying various individual components and weighted sums of the components. We estimated the dependency on the Violence History Subscale as a piecewise linear function, which we call $g_{\text{viol hist}}$. Next, in Figure 6, we plot the remainder after also subtracting this dependency on Violence History (that is, the remainder of the COMPAS violent score after subtracting both $f_{\text{viol age}}$ and $g_{\text{viol hist}}$) against the unweighted sum of the components of the History of Noncompliance Subscale, on which the violent score should also depend. There is not a sharp lower bound that is consistent across the horizontal axis, which means this sum, by itself, is not likely to be an additive term within COMPAS. Therefore, we do not estimate a dependency on the unweighted sum of items in the History of Noncompliance Subscale.

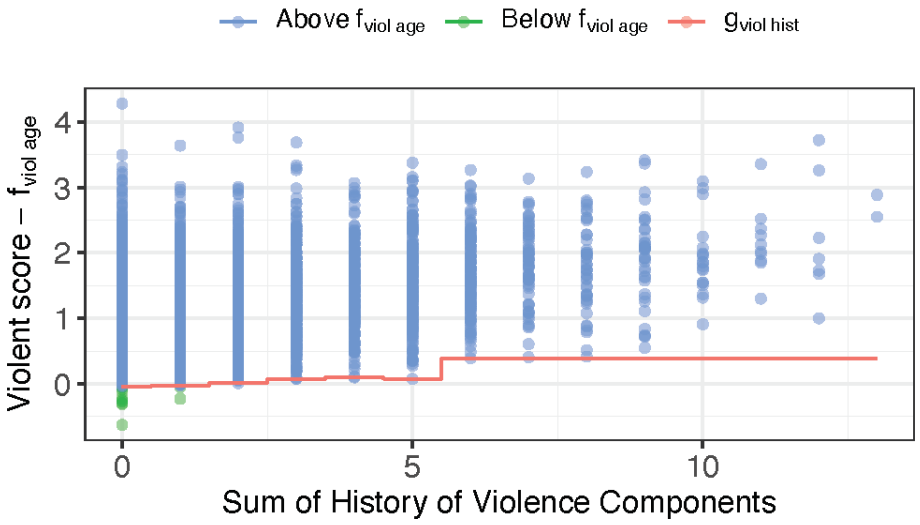


Figure 5. COMPAS violent score $-f_{\text{age}}$ vs. sum of History of Violence components. Green points are age outliers.

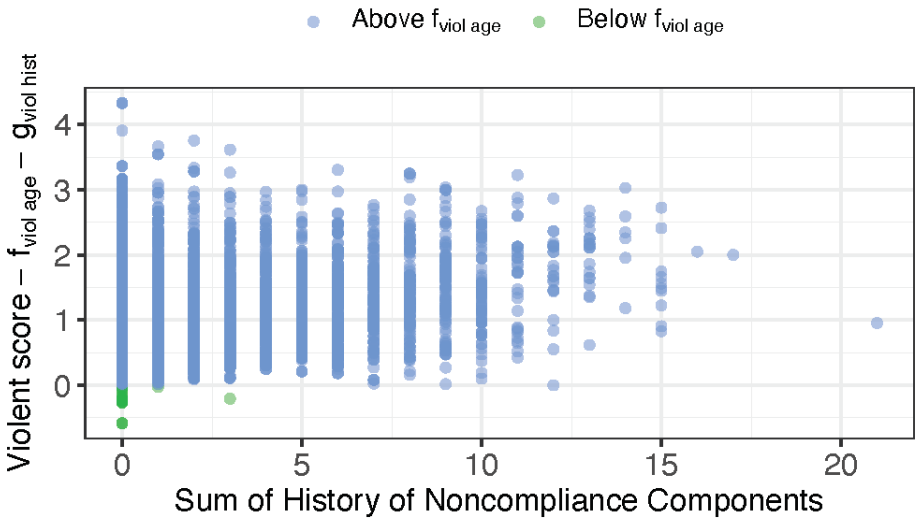
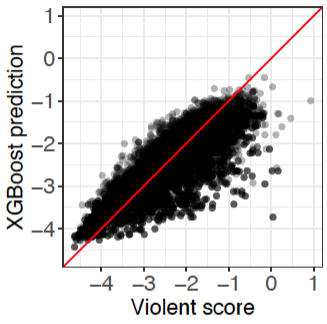


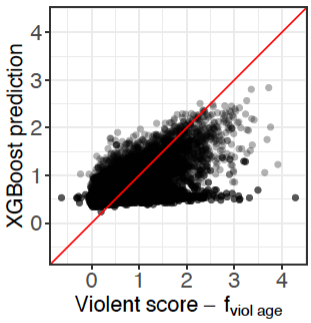
Figure 6. COMPAS violent score $-f_{\text{age}} - g_{\text{viol hist}}$ vs. sum of History of Noncompliance components. Green points are age outliers.

As with the COMPAS general score, we investigate if the COMPAS violent score depends on its subscale components in a more complex way. Figure 7 shows the results of three separate prediction problems using all of the components in the History of Violence and History of Noncompliance Subscales. From left to right, we use gradient boosted trees to predict the COMPAS violent score, the COMPAS violent score after subtracting $f_{\text{viol age}}$, and the COMPAS violent score after subtracting $f_{\text{viol age}}$ and $g_{\text{viol hist}}$. By comparing Figures 7(a) and 7(b), we observe that the predictions of the

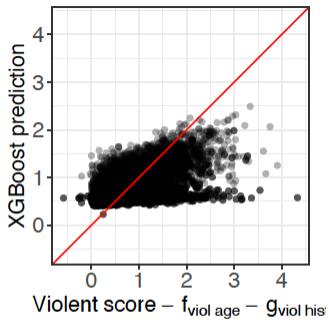
COMPAS violent score degrade substantially when $f_{\text{viol age}}$ is subtracted, indicating that $f_{\text{viol age}}$ makes a significant contribution to the COMPAS violent score. Predictions degrade far less between Figures 7(b) and 7(c)—thus, $g_{\text{viol hist}}$ seems to contribute little to the violent score. We were not able to see any dependence on the History of Noncompliance Subscale items. *We conclude that the dependence of the COMPAS violent score on criminal history, as captured by the Violence History and History of Noncompliance Subscales, seems to be weak.*



(a) Predictions of COMPAS violent score vs. actual values.



(b) Predictions of COMPAS violent score $-f_{\text{viol age}}$ vs. actual values.



(c) Predictions of COMPAS violent score $-f_{\text{viol age}}$ $-g_{\text{viol hist}}$ vs. actual values.

Figure 7. Predicting violent remainder.

2.5 Caveats

For both the general and violent COMPAS scores, we were unable to capture the remainder of the COMPAS score (i.e., after subtracting the reconstructed dependency on age) using the various criminal history variables that constitute the subscales.

There could be several reasons for this, including the following, among other things:

- It is possible that our data are flawed. We obtained these data from a combination of ProPublica, the Broward County Sheriff, as well as the Broward County Clerk's office. We believe most of these data should be approximately the same as the data entered into the COMPAS models. Furthermore, based on the analysis above, our age data on individuals seems to be high quality, so there is no a priori reason that the criminal history data would be substantially lower quality.

- It is also possible that the way we calculated the criminal history subscale items for COMPAS differs from the way the Broward County Sheriff's Office calculates them. Our data processing is discussed in the Appendix.

- It is possible that we did not hypothesize the correct model form used in COMPAS; that is, our machine learning models may not have been able to express the nonlinearities present in COMPAS. While this could be true, we used very flexible models that should be able to fit (or even overfit) the COMPAS scores. Thus, we believe this is not a likely explanation.

- It is possible that our data are incomplete. We know this is true, as COMPAS depends on factors other than criminal history. However, this leads to questions of what COMPAS can reasonably depend heavily on. Criminal history data are less noisy and less susceptible to manipulation than other survey questions; criminal history features do not depend on the survey-taker telling the truth about the answers. If COMPAS depended more heavily on survey questions than on criminal history, it could lead precisely to a kind of bias that we might want to avoid. For instance, if COMPAS did not depend heavily on the number of prior crimes, it might depend more heavily on socioeconomic proxies (e.g., "How hard is it for you to find a job ABOVE minimum wage compared to others?," which is one of several questions on the COMPAS questionnaire that directly relates to socioeconomic status).

- There is something in the procedure of calculating the COMPAS score that causes it to be calculated inconsistently. Since we do not know COMPAS, we cannot check this possibility. In the past, there have been documented cases where individuals have received incorrect COMPAS scores based on incorrect criminal history data (Wexler, 2017a, 2017b), and no mechanism to correct it after a decision was made based on that incorrect score. We do not know whether this happens often enough to influence the scatter plot in a visible way. However, this type of miscalculation is one of the biggest dangers in the use of proprietary models. As we know from the calculations of other scoring systems besides COMPAS, if the number of crimes is not properly taken into account, or if the scoring system is improperly calculated in other ways, it could lead (and has led in the past) to unfair denial of parole, and dangerous criminals being released pretrial.

Data-quality issues, either for us or for the COMPAS score itself, are not just a problem for our analysis, but a problem that is likely to plague almost every jurisdiction that uses COMPAS or any other secret model. This is further discussed in the next section.

2.6 ProPublica Seems to Be Incorrect in Its Claims of How COMPAS Depends on Race

If age and the components we have of the Criminal Involvement, History of Noncompliance, and History of Violence Subscales can explain COMPAS scores to only a limited extent, then either the few components of these subscales that we are missing, or the remaining subscales, Substance Abuse for

the violent score and Vocational/Educational for both scores, must be a large component of the scores. Reasoning these two subscales are highly correlated with race, we then tried to model the COMPAS remainders (i.e., after subtracting the age splines) with race as a feature, in addition to the available subscale components. Tables 4 and 5, respectively, show the results of several machine learning methods for predicting the general and violent score remainders. We see that these features cannot explain the COMPAS violent score remainders very well. *Thus, to conclude, we hypothesize that COMPAS has at most weak dependence on race* after conditioning on age and criminal history.

We replicated ProPublica's finding that a model with race predicts COMPAS well, but disagree with their conclusions. We repeated ProPublica's logistic regression on our slightly modified data set, leading to a model, provided in the Appendix in Table A3, whose race coefficient is large and significantly different from zero. Coefficients for age and race are both large.

There are several flaws in ProPublica's analysis. First, their modeling assumption for the logistic regression model is likely to be wrong, as we know from considering the age analysis above. They assumed a piecewise constant dependence on age, with three broad predefined age bins, which, according to our analysis, would be a poor approximation, particularly for younger ages. Second, the existence of an accurate model that depends on race is not sufficient to prove that COMPAS depends on race. Race is correlated with both criminal history and with age in this data set. Because ProPublica's dependence on age is probably wrong, it is easily possible that race would appear to be significant, regardless of whether COMPAS is actually using race or its proxies (aside from criminal history and age) as important variables. As shown in Tables 4 and 5, including race as a variable to predict COMPAS does not improve prediction accuracy. That is, for each model we created that incorporates race, we found another almost equally accurate model that does not incorporate race. Thus, it is not clear that race or its proxies (aside from criminal history and age) are necessarily important factors in COMPAS.

	Linear Model	**Random Forest**	**Boosting**	**SVM**
Without Race	0.573	0.533	0.521	0.524
With Race	0.562	0.524	0.506	0.514

Table 4. RMSE of Machine Learning Methods for Predicting COMPAS General Recidivism Raw Score after Subtracting f_{age} With and Without Race as a Feature.[7]

	Linear Model	**Random Forest**	**Boosting**	**SVM**

Without Race	0.472	0.460	0.452	0.462
With Race	0.464	0.447	0.443	0.447

Table 5. RMSE of Machine Learning Methods for Predicting COMPAS Violent Recidivism Raw Score after Subtracting $f_{\text{viol age}}$ With and Without Race as a Feature.[8]

In a separate analysis, Fisher, Rudin, and Dominici (2019) consider all models that approximate COMPAS with low loss, and among these, find models that depend the most and the least on race.

3. COMPAS Sometimes Labels Individuals With Long or Serious Criminal Histories as Low-Risk

We examine whether COMPAS scores can be low for individuals who pose a serious threat. Recently in California (Ho, 2017; Westervelt, 2017), a defendant with a long criminal history was released pre-trial after a criminal history variable was inadvertently mistyped into a scoring system as being much lower than its true value. The defendant murdered a bystander before his trial.

Typographical (data entry) errors are extremely common (Bushway, Owens, & Piehl, 2012), which means risk-score errors occur regularly. For instance, if each person's COMPAS questionnaire contains 100+ questions, even a 1% error rate could cause multiple wrong entries on almost every person's questionnaire. Data-entry errors are a serious problem for medical records (Wahi, Parks, Skeate, & Goldin, 2008), and in numerous other applications. The threat of typographical errors magnifies as the complexity of the scoring system increases; California had a very simple scoring system, and still typographical errors have caused serious events discussed above.

In what follows, we use publicly available records that can easily be found on the Internet. We have changed these names for our exposition.

Consider the case of Bryan Row, whose criminal history included trafficking cocaine and aggravated battery of a pregnant woman (felony battery—domestic battery by strangulation). He was given a COMPAS violent score of 1 (the lowest possible risk). A similar problem occurs for several individuals in the database. Table 6 shows several such individuals who have COMPAS scores that appear to have been calculated with incorrect inputs, or whose criminal history information somehow has not been considered within the COMPAS formula. None of these individuals have scores below the age spline (none are age outliers).

Name	COMPAS Violent Decile	# Arrests	# Charges	Selected Prior Charges	Selected Subsequent Charges
Shirley Darby	1	2	4	Aggravated Battery (F,1), Child Abuse (F,1), Resist Officer w/Violence (F,1)	
Joseph Salera	1	8	14	Battery on Law Enforc Officer (F,3), Aggravated Assault w/Dead Weapon (F,1), Aggravated Battery (F,1), Resist/obstruct Officer w/violence (F,1)	
Bart Sandell	1	9	15	Attempted Murder 1st Degree (F,1), Resist/obstruct Officer w/viol (F,1), Aggravated Battery Grt/Bod/Harm (F,1), Carrying Concealed Firearm (F,1)	Armed Sex Batt/vict, 12 Yrs + (F,2), Aggravated Assault w/Dead Weapon (F,3), Kidnapping (F,1)

Miguel Wilkins	1	11	22	Aggravated Battery w/Deadly Weapon (F,1), Driving Under the Influence (M,2), Carrying Concealed Firearm (F,1)	
Jonathan Gabbard	1	7	28	Robbery / Deadly Weapon (F,11), Possession Firearm Commission Felony (F,7)	
Brandon Jackel	1	22	40	Resist/obstruct Officer w/violence (F,3), Battery on Law Enforc Officer (F,2), Attempted Robbery Deadly Weapon (F,1), Robbery 1 / Deadly Weapon (F,1)	

Fernando Galarza	2	2	6	Murder in the First Degree (F,1), Aggravated Battery w/Deadly Weapon (F,1), Carrying Concealed Firearm (F,1)	
Nathan Keller	2	8	17	Aggravated Assault (F,5), Aggravated Assault w/Deadly Weapon (F,2), Shoot/Throw Into Vehicle (F,2), Battery Upon Detainee (F,1)	
Zachary Campanelli	2	11	21	Armed Trafficking In Cocaine (F,1), Possession Weapon Commission Felony (F,1), Carrying Concealed Firearm (F,1)	

Aaron Coleburn	2	16	25	Attempt Murder in the First Degree (F,1), Carrying Concealed Firearm (F,1), Felon in Possession of Firearm or Amm (F,1)	
Bruce Poblano	2	22	39	Aggravated Battery (F,3), Robbery / Deadly Weapon (F,3), Kidnapping (F,1), Carrying Concealed Firearm (F,2)	Grand Theft in the 3rd Degree (F,3)
Philip Sperry	3	11	16	Aggravated Assault w/Deadly Weapon (F,1), Burglary Damage Property>$1000 (F,1), Burglary Unoccupied Dwelling (F,1)	

Dylan Azzi	3	11	17	Aggravated Assault w/Deadly Weapon (F,2), Aggravated Assault w/Firearm (F,2), Discharge Firearm From Vehicle (F,1), Home Invasion Robbery (F,1)	Fail Register Vehicle (M,2)
Russell Michaels	3	9	23	Solicit to Commit Armed Robbery (F,1), Armed False Imprisonment (F,1), Home Invasion Robbery (F,1)	Driving While License Revoked (F,3)
Bradley Haddock	3	15	25	Attempt Sexual Battery / Victim 12+ (F,1), Resist/Obstruct Officer w/violence (F,1), Possession Firearm w/Alter/Remov Id# (F,1)	

Randy Walkman	3	24	36	Murder in the First Degree (F,1), Possession Firearm Commission Felony (F,1), Solicit to Commit Armed Robbery (F,1)	Petit Theft 100-300 (M,1)
Carol Hartman	4	5	16	Aggravated Battery w/Deadly Weapon (F,1), Felon in Possession of Firearm or Amm (F,4)	Resist/Obstruct w/o Violence (M,1), Possess Drug Paraphernalia (M,1)

Table 6. Individuals Whose COMPAS Violence Decile Score is Low (Low-Risk), but Who Have Significant Criminal Histories.[9]

While it is possible that COMPAS includes mitigating factors (employment, education, drug treatment) that reduce its score, it seems unlikely that they would reduce the score all the way to the lowest possible value, but since the COMPAS models are not published, we cannot actually determine this. According to Northpointe (2019), the only negatively weighted factors in COMPAS are age and age at first arrest, but according to our analysis above, these variables remain essentially constant with age for older individuals. This indicates there is no way to reduce a high score that might arise from a lengthy criminal history. Thus, what we are observing (long criminal histories with a low COMPAS violent score) should be impossible *unless inputs have been entered incorrectly or omitted from the COMPAS score altogether*.

COMPAS general or violent scores do not include the current charges. Thus, in the case of Juan Ortez in the ProPublica database, charged with a serious crime (kidnapping) but no prior crimes, he still received the lowest-risk COMPAS score of 1.

There are many individuals in the database whose COMPAS scores appear to be unreasonably high; however, it is possible that for those individuals, there are extra risk factors that cause them to be labeled high risk that are not in our database (e.g., incomplete criminal history information). Missing information would be able to explain COMPAS scores that seem too high, but it cannot explain COMPAS scores that are too low, such as the ones we presented above in Table 6. Figure 8 shows the predictions of a machine learning model versus COMPAS score. There are a huge number of individuals whose COMPAS score is much larger than the machine learning predictions, and also, there are many individuals for whom the machine learning model (a boosted decision tree) indicates high risk of recidivism, but the COMPAS score indicates a lower risk.

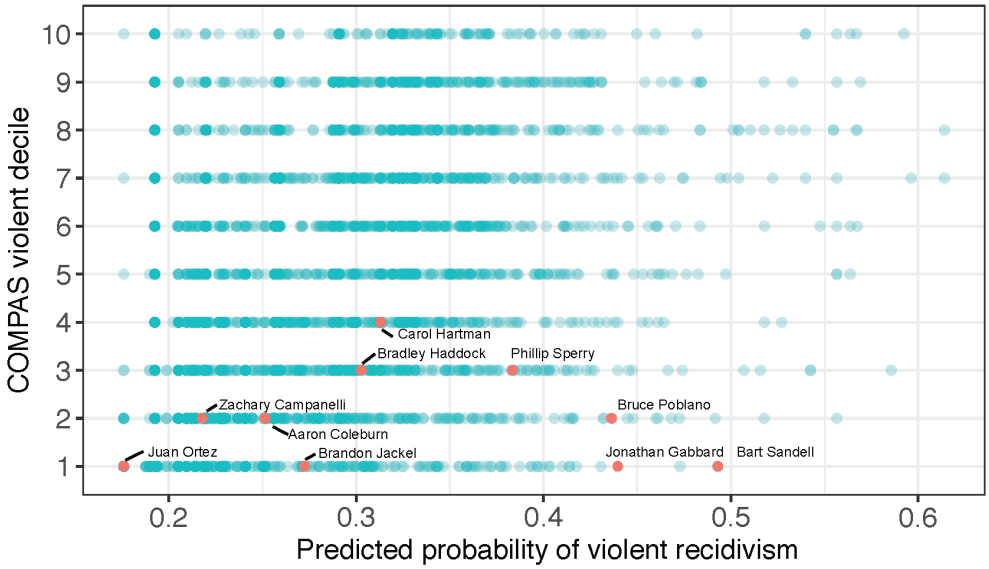


Figure 8. Predicted probability of violent recidivism vs. COMPAS violent decile score. Selected individuals listed in Table 6 are highlighted. Predicted probabilities were produced by a boosted tree model with 84.11% accuracy.

In cases like that of prisoner Glenn Rodrïguez, whose parole was denied because of a miscalculated COMPAS score (Wexler, 2017a, 2017b), he did not notice the error on his COMPAS form until after his parole was denied. Complicated forms, even if one is asked to check them over, lead to human error. We are certain, for instance, that there are still errors in this article, no matter how many times we have checked it over—and the longer the article, the more errors it probably contains.

Rodrïguez is a proven example of a data input error, but his case further demonstrates that *data input transparency alone is insufficient to help wronged defendants*—the scoring mechanism must also be transparent. Rodrïguez was unable to change his parole sentence because he was unable to demonstrate that the data error affected his COMPAS score.

4. Is Age Unfair? Fairness Through the Lens of Transparent Models

Differences in true/false positive rates do not consider other factors like age and criminal history. In ProPublica's regression analysis of the COMPAS score, ProPublica conditioned on age and criminal history among other factors, indicating *they thought COMPAS would hold to some notion of fairness had it depended only on age and criminal history*. (Otherwise, why condition on those two factors explicitly?) They used the significance of the coefficient on race to support their conclusion of bias against blacks. However, they used a piecewise constant term for age and did not handle unmeasured confounding, so this analysis was faulty. The faulty regression analysis leaves the differences in true/positive rates as their only remaining evidence of bias. However, the true/false positive rates are not conditioned on age and criminal history; that was why they performed the regression analysis, suggesting the regression could mitigate bias. In other words, ProPublica's second analysis (the regression model) was invalid because it assumed a particular dependence on age that appears to be false. Their first analysis (the true/false positive rate analysis) would also then have been invalid for *exactly the reasons why they conducted the second analysis* (which is that they would consider the model fair if COMPAS did not depend on race when conditioned on age and criminal history).

Age is a well-known determining risk factor for recidivism. Many recidivism scoring systems depend on age (Barnes & Hyatt, 2012; Helmus, Thornton, Hanson, & Babchinshin, 2012; Hoffman & Adelberg, 1980; Howard, Francis, Soothill, & Humphreys, 2009; Langton et al., 2007; Nafekh & Motiuk, 2002; Turner, Hess, & Jannetta, 2009; Zeng et al., 2017) since it has no direct causal relationship with race (race does not cause age, age does not cause race), and it is a good predictor of future crime. For adults in Broward County, the risk of recidivism decreases with age.[10]

On the other hand, in the Broward County data, African American people tend to be disproportionately represented at younger ages than Caucasians; the median age of a COMPAS assessment on an African American person is 27 years whereas the median age of a Caucasian is 33 years.[11] This means that more African Americans will be labeled as high risk than Caucasians. This also means that more African Americans will be *mistakenly* labeled as high risk than Caucasians. It also means that more Caucasians will be mistakenly labeled as low risk than African Americans. Because of the difference in ages between Blacks and Whites in the data set, even models that consider only age and criminal history can be as 'unfair' as COMPAS by ProPublica's true/false positive rate analysis.

Figure 9 shows the true positive rate (TPR), false positive rate (FPR), true negative rate (TNR) and false negative rates (FNR) for the model *age*, which is defined to be "If age $\leq$ 24, then predict arrest within 2 years, otherwise predict no arrest."

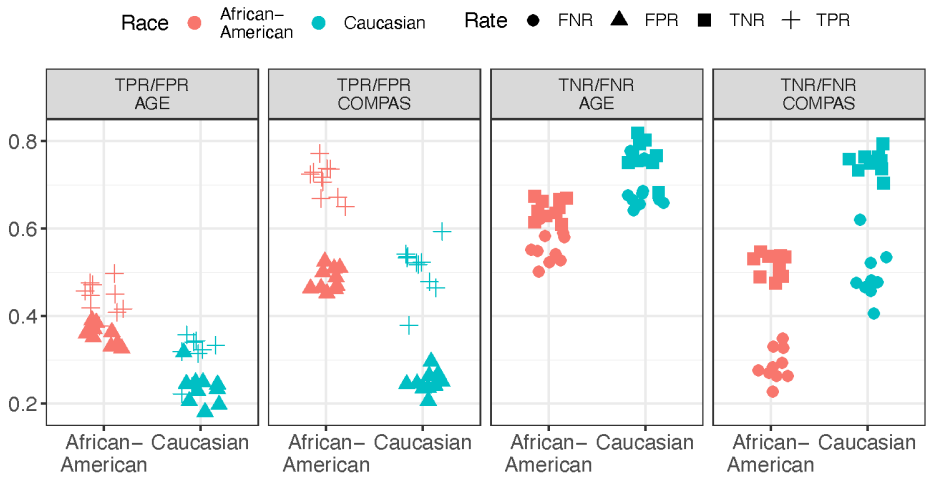


Figure 9. Rates for the simple age model and for the COMPAS score. *Age* appears also to be unfair.

The figure also shows the rates for the COMPAS general recidivism score for comparison. The data were divided into 10 randomly chosen folds, and the rates are plotted for all folds, showing a consistent pattern across folds. Indeed, for the age model, we observe higher false positive rates for African Americans, and higher false negative rates for Caucasians. There is an elevated ≈10% higher FPR for African Americans than for Caucasians for *age*, and an ≈10% higher FNR for Caucasians than African Americans for age. These differing levels constitute one of the accusations of unfairness by ProPublica, which means that *age* is an unfair risk prediction model by this definition. COMPAS seems to be more 'unfair' than *age*, but as we have seen, it may be possible to explain this unfairness by a combination of age and other features that differ between the distributions of African Americans and Caucasians and have little to do with the COMPAS score itself. In fact, we also know from Angelino et al. (2018) that a very simple model involving age and the number of priors is just as unfair as COMPAS by this definition.

The many definitions of fairness often contradict each other. Had ProPublica considered just the transparent model *age* (instead of analyzing COMPAS), it would have been easy to see that the difference in age populations between African Americans and Caucasians caused their two notions of fairness to disagree. In that case, would they still have written an article claiming that the difference in true and false positive rates meant that COMPAS was unfair? Or would they have claimed that the use of age was unfair rather than conditioning on it? Or would they have claimed the data collection process was unfair? Using a transparent model can sometimes enlighten an understanding of the debate on fairness.

The consequences of using age in criminal risk assessments are explained nicely by Patrick (2018); however, the use of criminal history to assess fairness is confusing for additional reasons. If we do use

criminal history in risk prediction, since African Americans tend to have longer criminal histories, their scores will be worse. On the other hand, if we do not use criminal history, our risk predictions would be worse. In that case, we could be releasing dangerous criminals based on poor pretrial risk assessments, which leads to poor decisions for the public.

Of course, if we eliminate the most important predictors of recidivism (age and criminal history) on grounds of leading to unfair models, it is not clear that any useful predictors of criminal recidivism remain. In that case, we could be stuck in the situation where a human decision maker provides nontransparent, potentially biased decisions, without the help of statistics calculated on databases.

The point of this exercise is not to determine whether the age model is fair by any given definition: the age model is transparent, which makes it much easier to debate, and useful for explaining different possible definitions of fairness and how they may never intersect. ProPublica's regression analysis seems to assume that using age in a risk prediction model is reasonable. But *is* age unfair? If we cannot decide on whether the age model is fair, we certainly cannot decide on whether COMPAS is unfair. However, it is certainly much easier to debate about the transparent and simple age model than about a black box scoring system. While a review of the numerous definitions of fairness (Berk, Heidari, Jabbari, Kearns, & Roth, 2018; Fairness, Accountability, and Transparency in Machine Learning [FATML], 2018) is outside the scope of this work, a potentially easy way to alter the definition of fairness is to control for nonprotected covariates such as age. In that case, as long as predictions for African Americans and Caucasians have equal true/false positive rates for each age group, then the model would be considered fair. Of course, a problem with this definition is that any policy that targets young people disproportionately affects African Americans.

5. Discussion

After attempting to isolate COMPAS's dependence on age, we investigated how much COMPAS can depend on criminal history and proxies for race. We found that it is unlikely that COMPAS depends heavily on either of them. Machine learning methods for predicting COMPAS scores performed equally well with or without direct knowledge of race. This seems to contradict ProPublica's claims, but ProPublica's methodological assumptions (at least about COMPAS's dependence on age) were wrong, which caused their conclusions to be faulty.

Northpointe claims the current charge is not helpful for prediction of future violent offenses (Nortpointe, 2019). (Oddly, they have a separate "Current Violence" scale that includes the current charges, but that is not claimed to be predictive.) How much should one weigh the current charges with the COMPAS scores? This is not clear. Because COMPAS is a black box, it is difficult for practitioners to combine the current charge (or any other outside information) with the COMPAS scores. Because the current charges are separate, COMPAS scores are not single numbers that

represent risk. Instead their interpretation entails more degrees of freedom. Could decision makers fail to realize that the COMPAS score does not include the current charge? Perhaps this alone could lead to faulty decision-making.

We showed examples where COMPAS scores can label individuals with long criminal histories as low risk. This could easily stem from a lack of transparency in COMPAS and could lead to dangerous situations for the public. Even if COMPAS were completely fair by some reasonable definition of fairness, this would not stop it from being miscalculated, leading to a form of procedural unfairness. Since it is known that COMPAS is no more useful for predicting recidivism than simple, interpretable models, there is no good reason to continue using complicated, expensive, error-prone proprietary models for this purpose. There is a mystique behind the use of black box models for prediction. However, just because a model is a proprietary does not mean it is superior to a publicly available model (Rudin, 2019; Rudin & Radin, 2019).

Interestingly, a system that relies only on judges—and does not use machine learning at all—has similar disadvantages to COMPAS; the thought processes of judges is (like COMPAS) a black box that provides inconsistent error-prone decisions. Removing COMPAS from the criminal justice system, without a transparent alternative, would still leave us with a black box.

Privacy of data should be more of a concern than it presently is. If COMPAS does not depend heavily on most of the 137 variables, including the proxies for socioeconomic status, it is not clear if Northpointe is justified in collecting such private information. COMPAS is a risk *and needs* assessment, but is substantial private information necessary to assess an individual's needs? All evidence suggests it does not seem to be necessary for estimating risk. Determination of needs seems to be a complicated causal question about who benefits from what types of treatments. This issue is beyond the scope of this article, but is important. Northpointe's control over criminal risk scores is analogous to Equifax's control over credit scores, and leads to inherent privacy risks.

Thus, our findings indicate that some form of unfairness caused by COMPAS can affect almost everyone involved in the justice system: 1) Lack of transparency makes it difficult to assess any of the myriad forms of fairness, leading to faulty arguments like those of ProPublica. (*Lack of transparency can hide bias toward underrepresented groups, or conversely, it can make fair models seem biased.*) 2) The unnecessary complexity of COMPAS could cause injustice to those hurt by typos in the COMPAS computation, and as a result, were given extra long sentences or denial of parole. (*This is procedural unfairness.*) Further, typos can lead to dangerous individuals being released, which becomes a greater risk with complicated models. 3) It is possibly unfair to the taxpayers and judicial system to pay for the collection of long COMPAS surveys and COMPAS predictions when simpler, transparent, options are available. (*Poor designation of public resources is unfair to everyone.*) 4) The subgroup of people who provided very private personal information (e.g., about their family history of crime or poverty) to

Northpointe has potentially been wronged. (*There is a form of privacy unfairness in being forced to provide personal information to an entity when it is unnecessary to do so.*)

The problems with COMPAS pertain to many industries. Without community standards or policy requirements for transparency, business considerations disincentivize creators of models to disclose their formulas. However, this lack of transparency is precisely what allows errors to propagate and results in damage to society (Rudin, 2019). Merely being able to explain black box models is not sufficient to resolve this—the models need to be fully transparent, and in criminal justice, there is no evidence of a loss in predictive accuracy for using a transparent model.

Data Repository/Code

Our code is here: https://github.com/beauCoker/age_of_unfairness

Disclosure Statement

This study was partially supported by Arnold Ventures.

Acknowledgments

We offer thanks to the Broward County Sheriff's office.

Appendix

Supporting Figures and Tables

Probability of Recidivism as a Function of Age

Figure A1 shows the probability of a new charge within two years as a function of age. The probability is a decreasing function of age.

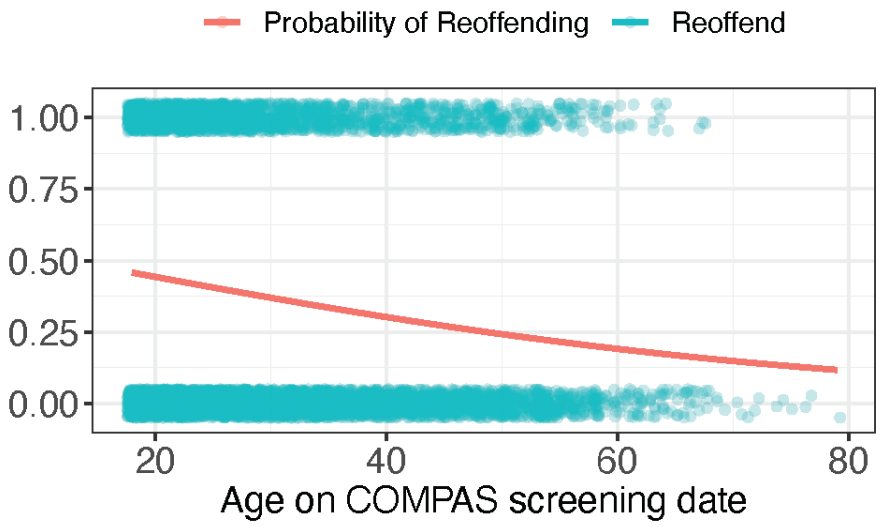


Figure A1. Probability of a new charge within two years as a function of age (red). The blue scatter plot is useful for understanding the distribution of ages of individuals; each individual who was arrested has a dot at their age on the horizontal axis, and a "1" on the vertical axis

Age Histograms

Figure A2 shows the normalized histograms of African Americans and Caucasians within Broward County who were evaluated via COMPAS between the beginning of 2013 and the end of 2014. These histograms do not involve COMPAS scores themselves, only information about the set of individuals who received COMPAS scores. The histogram for African Americans is skewed to the left, which means African Americans tend to be younger on average when their COMPAS score is calculated in the Broward County data set.

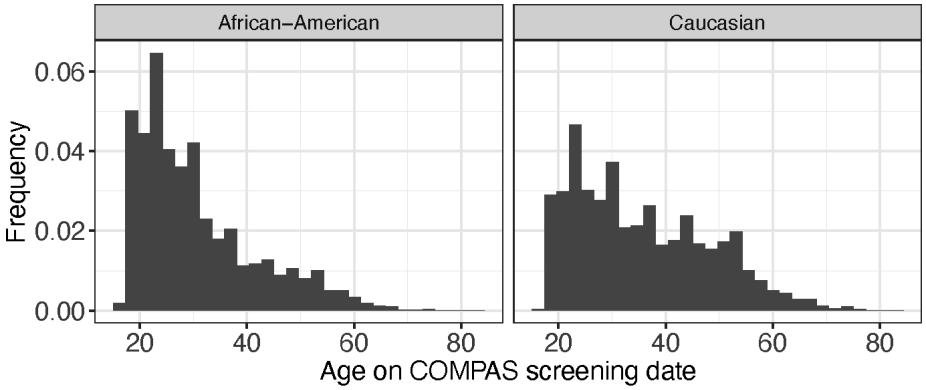


Figure A2. Normalized histograms of age for African Americans and Caucasians within the Broward County data set.

Predictions of Recidivism With and Without Race

Tables A1 and A2 show predictions of recidivism and violent recidivism, respectively, with and without race as a feature. The results are very similar with and without race.

Table A1. Misclassification Error of Machine Learning Methods for Predicting General Recidivism With and Without Race as a Feature.[12]

	Logistic Regression	Random Forest	Boosting	SVM
Without Race	0.330	0.321	0.312	0.306
With Race	0.328	0.318	0.313	0.313

Table A2. Misclassification Error of Machine Learning Methods for Predicting Violent Recidivism With and Without Race as a Feature.[13]

	Logistic Regression	Random Forest	Boosting	SVM
Without Race	0.159	0.165	0.156	0.158
With Race	0.159	0.161	0.155	0.160

Fitting f_{age} and $f_{\text{viol age}}$

Checking Data Assumptions

As explained in Section 2.2, we make the assumption that there exist individuals in our data with the lowest possible COMPAS score for each age. We do not have all the inputs to the COMPAS score, but if the assumption holds for the inputs we have access to, it lends evidence that the Data Assumption holds generally. In Figure A3, we show the counts of individuals who have current age equal to age-at-first-arrest and have all zero values for the COMPAS subscale inputs within our data.

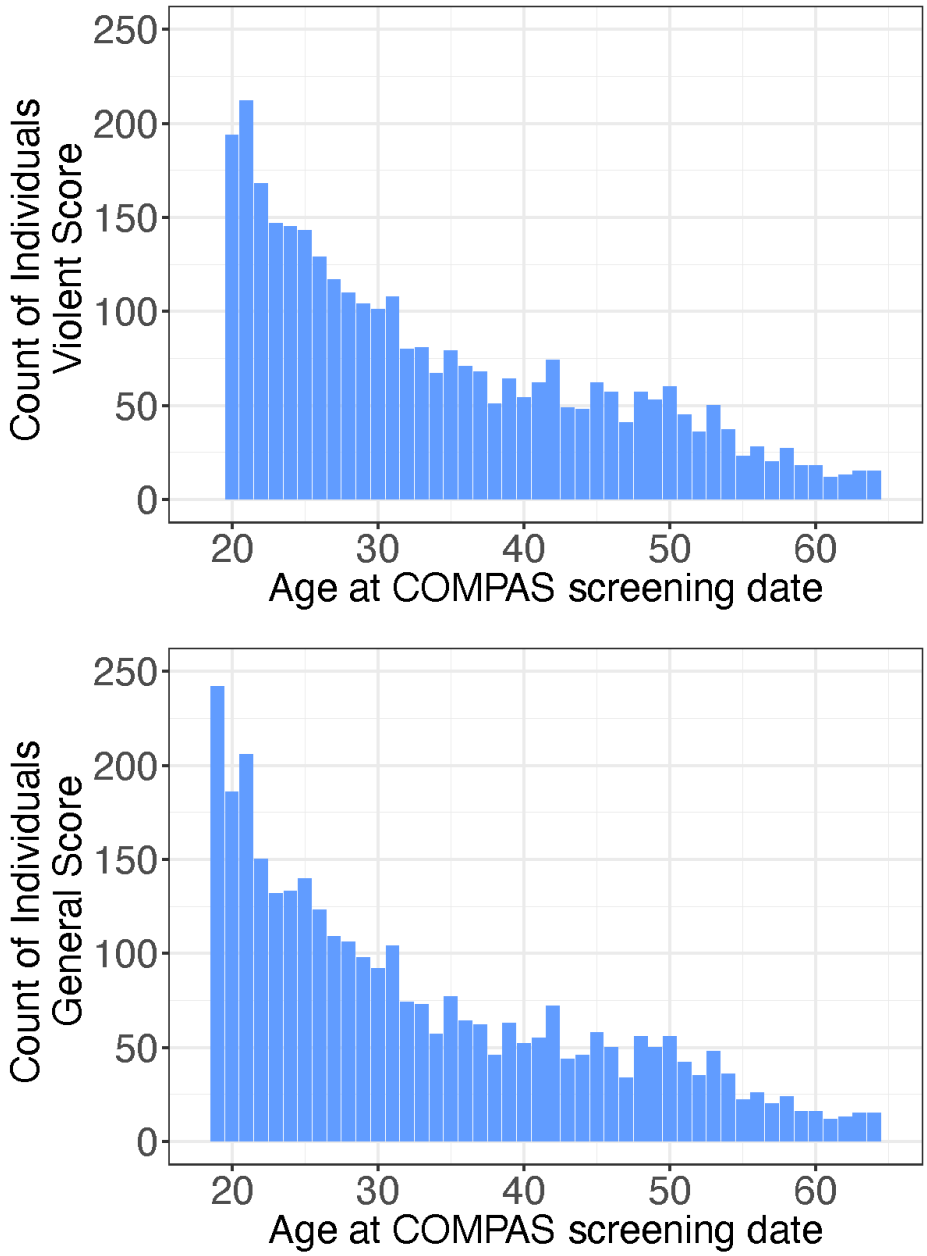


Figure A3. Age vs: the count of individuals who satisfy two conditions: (1) Age equals age-at-first-arrest, (2) All zeros for the subscale inputs in our data that correspond to the particular COMPAS score we consider (i.e., Criminal Involvement Subscale for the general score; History of Violence Subscale and History of Noncompliance Subscale for the violent score).

Age-at-First-Arrest

The COMPAS lower bounds f_{age} and $f_{\text{viol age}}$ are defined by the current age at which the COMPAS score is computed, not the age-at-first-arrest. We chose to begin the analysis with current age because (1) the relationship between current age and the COMPAS score is the clearest in the data, and (2) according to the COMPAS documentation, the age variables are the only variables that have a linear

relationship with the score. Figure A4 shows that there is not a smooth lower bound for COMPAS vs. age-at-first-arrest as there was with the current age.

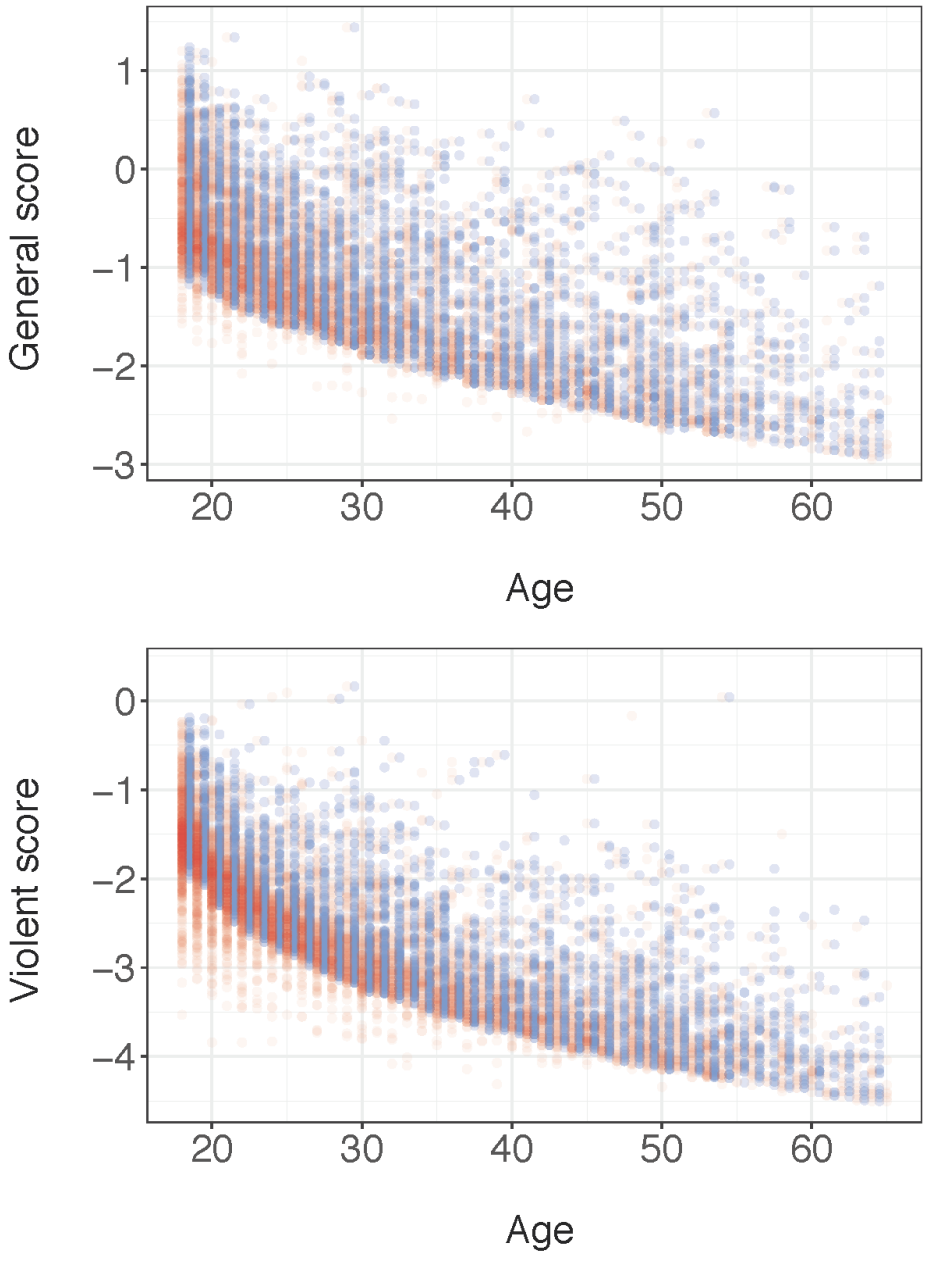


Figure A4. COMPAS raw score vs. the age variables. age-at-first-arrest is depicted in red, while current age is depicted in blue and jittered to the right. There is no clear pattern for the lower bound with age-at-first-arrest, as there is for current age. Thus, age-at-first-arrest is not used to define an initial lower bound.

Individuals on the Lower Bound Having the Lowest Possible COMPAS Scores for Their Age

The functions f_{age} and $f_{\text{viol age}}$ are defined by individuals who have zero values for the subscale components we can compute, and lie close to the age spline (i.e., were not deemed age outliers)—in other words, the individuals who *could* satisfy the Data Assumption. While there is no way for us to know if these individuals truly satisfy the Data Assumption, in this section we explain why we believe these individuals actually do satisfy the assumption.

Figure A5 shows raw COMPAS scores for these individuals who define the age splines. For many ages, there are several individuals whose COMPAS raw scores have identical values. The number of such individuals for each age is shown by the size of the ball in the figure. Their current age is usually equal to their age-at-first-arrest. Figure A6 shows this, where individuals with current age equal to age-at-first-arrest are plotted in blue, and others are in red.

It is possible that these individuals have nonzero values for the unobserved subscale components, which would imply that the true age functions could lie below f_{age} and $f_{\text{viol age}}$. However, we believe that Figure A5—which shows that for many age values there are multiple individuals with exactly the same COMPAS raw score on the lower bound of our data—combined with the smoothness of the lower bound, provides compelling evidence that all these individuals actually have zero values for the unobserved subscale components. Some alternative hypotheses and the reasons why we find them unlikely, are outlined here:

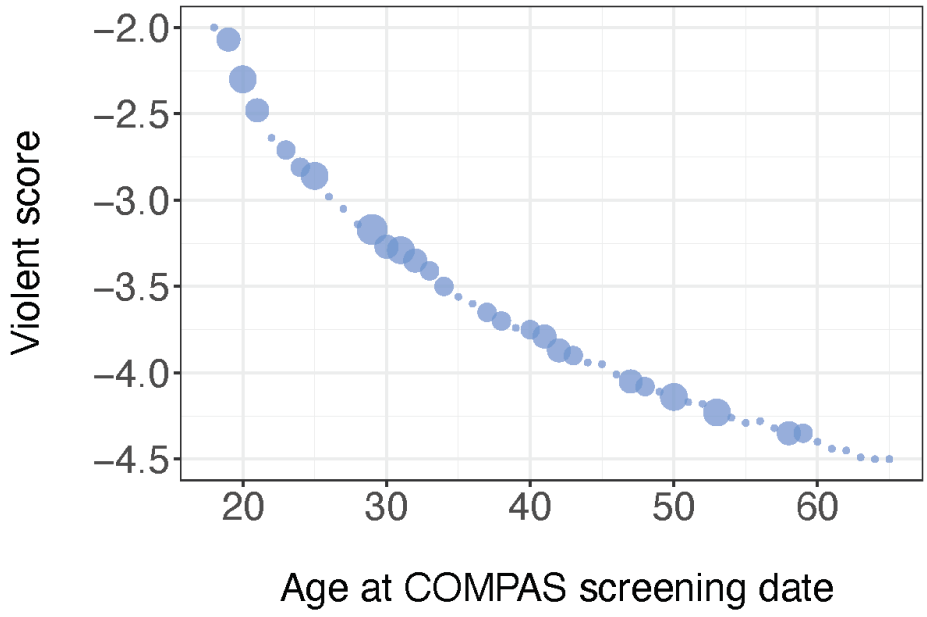


Figure A5a. Lowest violent COMPAS raw score for each age, for individuals with no criminal history and no history of noncompliance, excluding age outliers. The smallest balls represent only one individual; the larger balls represent 7 individuals with identical COMPAS raw scores. 83 individuals are represented in this plot.

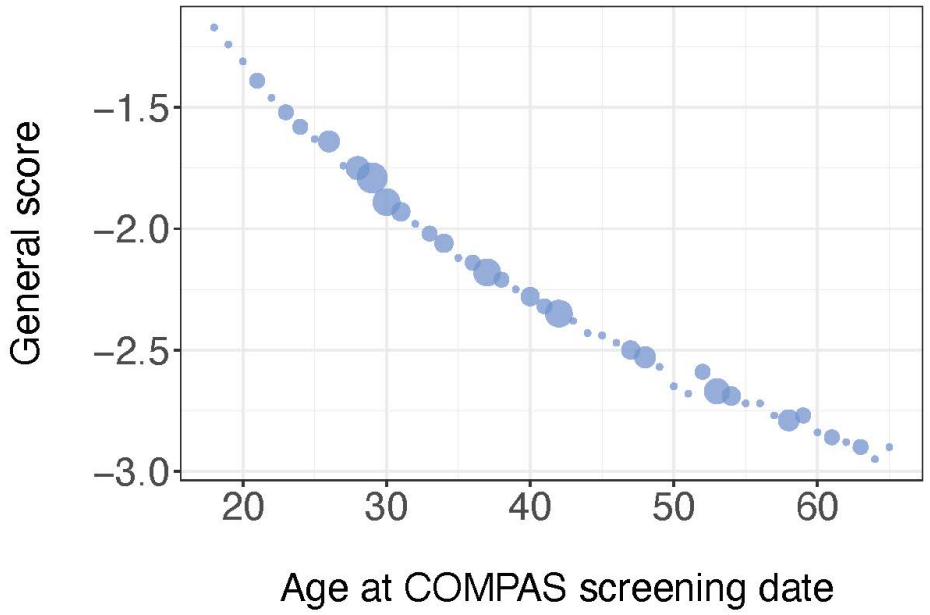
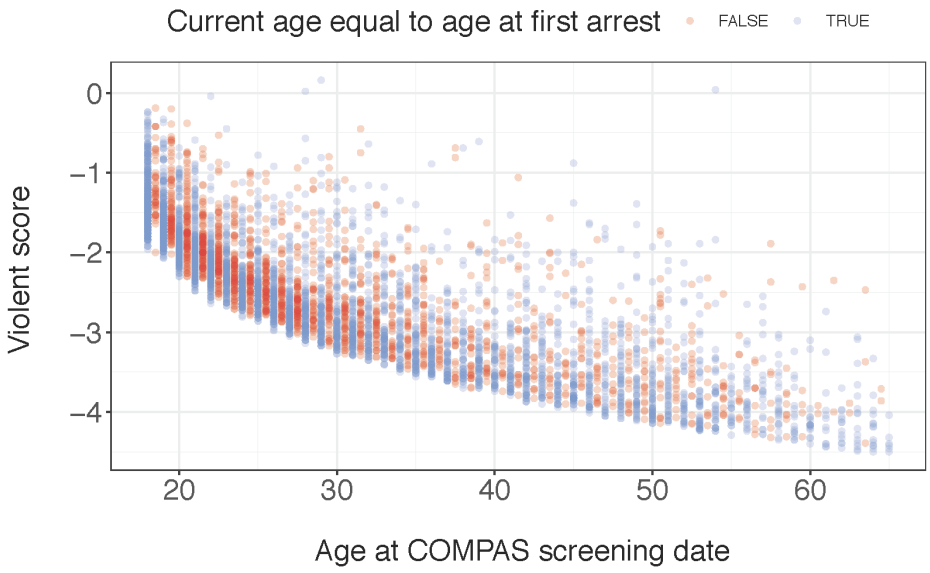


Figure A5b. Lowest general COMPAS raw score for each age, for individuals with no criminal involvement, excluding age outliers. The smallest balls represent only one individual; the larger balls represent 6 individuals with identical COMPAS raw scores. 103 individuals are represented in this plot.



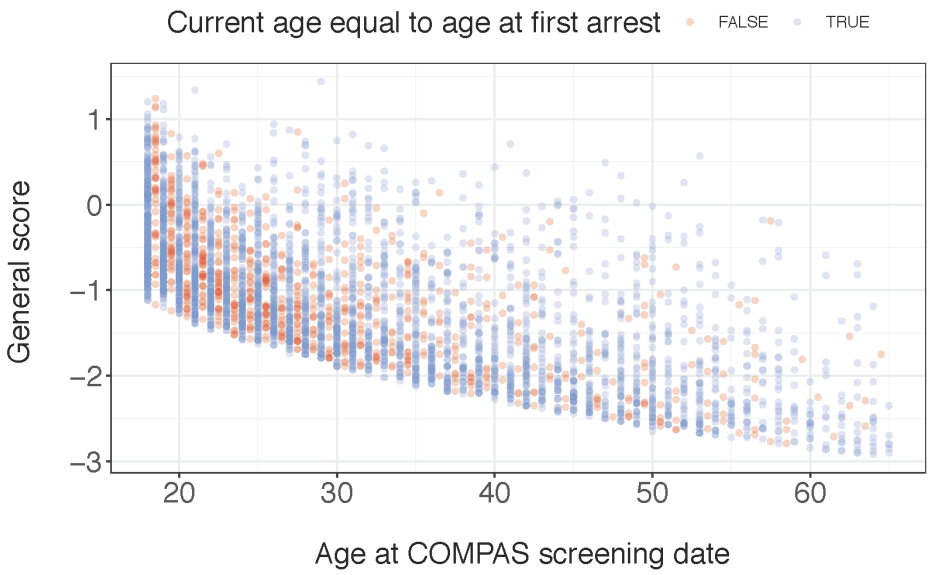


Figure A6. COMPAS raw score vs. current age. Individuals such that current age is equal to age-at-first-arrest are shown in blue, whereas others for which current age is not equal to age-at-first-arrest are shown in red. We define the lower boundary for the Violent Score using individuals that have the additional constraint that they cannot have violence history or noncompliance history; for the General Score, we use the constraint that individuals cannot have criminal involvement history. Note that individuals on the lower bound are mostly blue—that is, individuals have age equal to age at first arrest. Others tend to have higher COMPAS scores.

- Some individuals on the lower bound have zeros for unobserved subscale components, while others have nonzero values for the unobserved subscale components, which would mean that the true lower bound is lower than the one we found. We find this unlikely for two reasons. Since age and age-at-first-arrest are the only inputs with negative contribution to the COMPAS score, if we fix age, COMPAS scores for the individuals with positive values for unobserved subscale components should be higher than the lowest possible COMPAS score. Thus, for these individuals to lie on the lower bound we observe, there cannot be *any* individuals in the data who have the lowest possible COMPAS score for that age, aside from the few age outliers. Second, this would have to occur in a systematic way, to create the smooth, nonlinear lower bound we observe.

- All individuals on the lower bound have nonzero values for unobserved inputs because some of these inputs contribute negatively to the COMPAS score, thereby negating the positive contribution from other inputs, allowing the individual to have the lowest possible COMPAS score. We find this unlikely because, according to the COMPAS documentation, age and age-at-first-arrest are the only variables that contribute negatively to the COMPAS score. As discussed above, age-at-first-arrest equals age for most of the individuals on the lower bound, so there is no contribution from age-at-first-arrest. Though we claim the COMPAS documentation is not completely accurate, we believe it is reliable in this respect because it would be unintuitive for the other inputs to contribute negatively to the COMPAS scores.

- Some individuals on the lower bound have nonzero values for unobserved inputs, but manual overrides occurred for all of these individuals so that the only contribution to their score was age. This is possible, but implies that the only factor influencing their COMPAS scores was age, which then agrees with our hypothesis.

For these reasons, we believe that these individuals have the lowest possible COMPAS scores for their age and lie on the true age functions.

An Alternative Explanation for the Nonlinearity in Age

In this section, we briefly consider an alternative explanation for the nonlinearity in age. As observed in Section 4, there are a greater proportion of individuals at lower ages than at higher ages in our data. If we consider the possibility that age follows a Gaussian distribution with respect to the COMPAS score, it is possible that the nonlinear behavior we observe in the lower bound of individuals' COMPAS raw scores vs current age is induced by more extreme values appearing when we have more samples. However, exploratory analysis mitigates this concern.

Figure A7 shows the number of individuals per age group. As Figure A8 shows, sampling min (150, number of individuals per age) for each age group does not alter the nonlinearity of the lower bound. Moreover, polynomials fit using only the sampled data are very close to the polynomials fit using the full data.

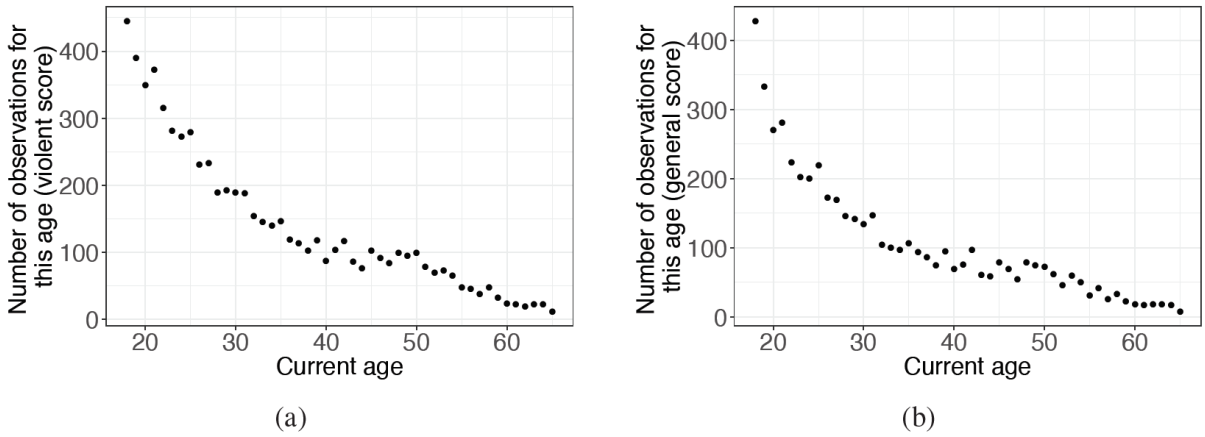


Figure A7. Number of observations for each age.

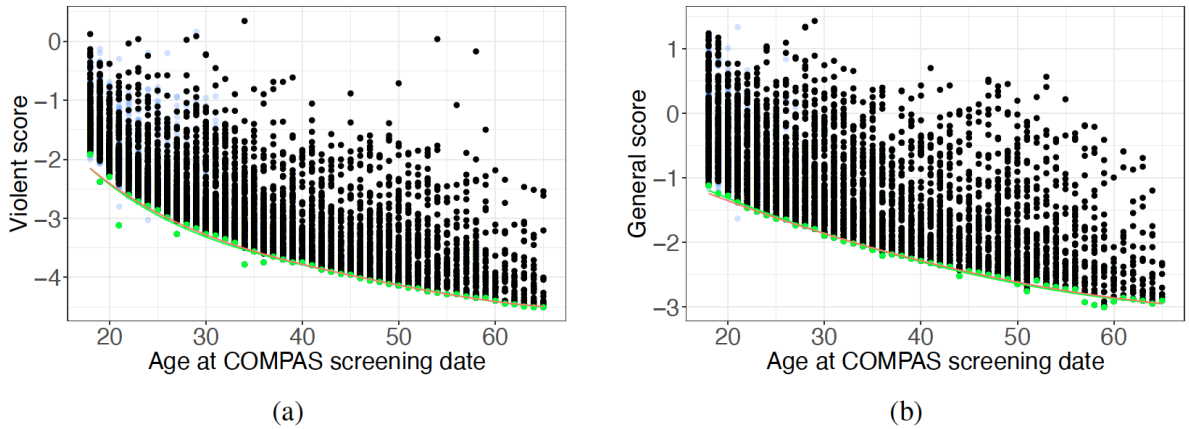


Figure A8. COMPAS raw scores versus age. Polynomials fitted using the complete data are in red. The polynomials fitted using a sample of n = min(150; number individuals per age) are in green. The two bounds are extremely close, mitigating the concern that the nonlinearity is due to the number of observations per age group.

Logistic Regression

We attempt to replicate ProPublica's logistic regression of the COMPAS score category (Medium or High Risk vs. Low Risk) on various features, including race. Coefficient estimates and standard errors are shown in Table A3. Since recidivism (the outcome for our recidivism prediction models) is used as a covariate in ProPublica's analysis, we exclude any observation for which there is less than two years of data beyond the screening date. Note that if two-year recidivism is used in ProPublica's model, it is using information that by definition is not available at the time that the COMPAS score is calculated.

Table A3. Logistic Regression Coefficient Estimates and Standard Errors Computed in a Similar Way to ProPublica.[14]

	Our Results		**ProPublica's Results**	
	Estimate	**Standard Error**	**Estimate**	**Standard Error**
Female	0.124	0.085	0.221***	0.080
Age: Greater than 45	-1.489***	0.130	-1.356***	0.099
Age: Less than 25	1.445***	0.071	1.308***	0.076
Black	0.521***	0.073	0.477***	0.069
Asian	-0.272	0.504	-0.254***	0.478

Hispanic	-0.301*	0.130	-0.428***	0.128
Native American	0.391	0.678	1.394*	0.766
Other	-0.714***	0.160	-0.826***	0.162
Number of Priors	0.155***	0.007	0.269***	0.011
Misdemeanor	-0.464***	0.070	-0.311***	0.067
Two year Recidivism	0.492***	0.069	0.686***	0.064
Constant	-1.593***	0.082	-1.526***	0.079

Data Processing

Our data includes the same raw data collected by ProPublica, which includes COMPAS scores for all individuals who were scored in 2013 and 2014, obtained from the Broward County Sheriff's Office. There are 18,610 individuals, but we follow ProPublica in examining only the 11,757 of these records which were assessed at the pretrial stage. We also used public criminal records from the Broward County Clerk's Office to obtain the events/documents and disposition for each case, which we used in our analysis to infer probation events.

ProPublica processed the raw data (Angwin et al., 2016; Larson et al., 2016), which includes charge, arrest, prison, and jail information, into features aggregated by person, like the number of priors or whether or not a new charge occurred within two years. While ProPublica published the code for their analysis and the raw data, they did not publish the code for processing the raw data. Thus we did that from scratch.

The *screening date* is the date on which the COMPAS score was calculated.

- Our features correspond to an individual on a particular screening date. If a person has multiple screening dates, we compute the features for each screening date, such that the set of events for calculating features for earlier screening dates is included in the set of events for later screening dates.

- On occasion, an individual will have multiple COMPAS scores calculated on the same date. There appears to be no information distinguishing these scores other than their identification number. We take the scores with the larger identification number.

- Any charge with degree (0) seems to be a very minor offense, so we exclude these charges. All other charge degrees are included, meaning charge degrees other than felonies and misdemeanors are included.

- Some components of the Violence Subscale require classifying the type of each offense (e.g., whether or not it is a weapons offense). We infer this from the statute number, most of which correspond to statute numbers from the Florida state crime code.

- The raw data includes arrest data as well as charge data. Because the arrest data does not include the statute, which is necessary for the Violence Subscale, we use the charge data and not the arrest data throughout the analysis. While the COMPAS subscales appear to be based on arrest data, we believe the charge data should provide similar results.

- For each person on each COMPAS screening date, we identify the offense—which we call the *current offense*—that we believe triggered the COMPAS screening. The *current offense date* is the date of the most recent charge that occurred on or before the COMPAS screening date. Any charge that occurred on the current offense date is part of the current offense. In some cases, there is no prior charge that occurred near the COMPAS screening date, suggesting charges may be missing from the data set. For this reason we consider charges that occurred within 30 days of the screening date for computing the current offense. If there are no charges in this range, we say the current offense is missing. For any part of our analysis that requires criminal history, we exclude observations with missing current offenses. All components of the COMPAS subscales that we compute are based on data that occurred prior to (not including) the current offense date, which is consistent with how the COMPAS score is calculated according to Northpointe (2019).

- The events/documents data includes a number of events (e.g., File Affidavit Of Defense or File Order Dismissing Appeal) related to each case, and thus to each person. To determine how many prior offenses occurred while on probation, or if the current offense occurred while on probation, we define a list of event descriptions indicating that an individual was taken on or off probation. Unfortunately, there appear to be missing events, as individuals often have consecutive On or consecutive Off events (e.g., two “On" events in a row, without an “Off" in between). In these cases, or if the first event is an Off event or the last event is an On event, we define two thresholds, t_{on} and t_{off}. If an offense occurred within t_{on} days after an On event or t_{off} days before an Off event, we count the offense as occurring while on probation. We set t_{on} to 365 and t_{off} to 30. On the other hand, the number of times on probation feature is just a count of On events and the number of times the

probation was revoked feature is just a count of File Order of Revocation of Probation event descriptions (i.e., there is no logic for inferring missing probation events for these two features).

- Age is defined as the age in years, rounded down to the nearest integer, on the COMPAS screening date.

- Recidivism is defined as any charge that occurred within two years of the COMPAS screening date. For any part of our analysis that requires recidivism, we use only observations for which we have two years of subsequent data.

- A juvenile charge is defined as an offense that occurred prior to the defendant's 18th birthday.

Machine Learning Implementation

Here we discuss the implementation of the various machine learning methods used in this paper. To predict the COMPAS general and violent raw score remainders (Tables 2, 4, and 5), we use a linear regression (base `R`), random forests (`randomForest` package), Extreme Gradient Boosting (`xgboost` package), and SVM (`e1071` package). To clarify, we predict the COMPAS raw scores (not the decile scores, since these are computed by comparing the raw scores to a normalization group) after subtracting the age polynomials (f_{age} for the general raw score and $f_{\text{viol age}}$ for the violent raw score). For XGBoost and SVM we select hyperparameters by performing 5-fold cross validation on a grid of hyperparameters and then retrain the method on the set of hyperparameters with the smallest cross validation error. For random forest we use the default selection of hyperparameters. For the COMPAS general raw score remainder, we use the available Criminal Involvement Subscale features (Table A6), while for the COMPAS violent raw score remainder, we use the available History of Violence Subscale and History of Noncompliance Subscale features listed in tables Tables A4 and A5, respectively. For both types of COMPAS raw scores, we also use the age-at-first-arrest. Race and age at screening date may or may not be included as features, as indicated when the results are discussed. To predict general and violent 2-year recidivism (Tables A1 and A2), we use the same methods, features, and cross validation technique as used to predict the raw COMPAS score remainders, except we adapt each method for classification instead of regression (for linear regression, we substitute logistic regression) and we include the current offense in the features. All code is written in `R` and is available on GitHub[15].

Subscale Tables

The features that compose the subscales used by COMPAS and that we use for prediction are listed in Tables A4-A8. The Criminal History, Substance Abuse, and Vocation/Education Subscales (Tables A6, A8, and A7, respectively) are inputs to the COMPAS general recidivism score, while the History of

Violence, History of Noncompliance, and Vocation/Education Subscales (Tables A4, A5, and A7, respectively) are inputs to the COMPAS violent recidivism score.

Table A4. History of Violence Subscale.[16]

Subscale Items	Values
Prior juvenile felony offense arrests	**0, 1, 2+**
Prior violent felony property offense arrests	**0, 1, 2, 3, 4, 5+**
Prior murder/voluntary manslaughter arrests	**0, 1, 2, 3+**
Prior felony assault offense arrests (excluding murder, sex, or domestic violence)	**0, 1, 2, 3+**
Prior misdemeanor assault offense arrests (excluding murder, sex, or domestic violence)	**0, 1, 2, 3+**
Prior family violence arrests	**0, 1, 2, 3+**
Prior sex offense arrests	**0, 1, 2, 3+**
Prior weapons offense arrest	**0, 1, 2, 3+**
Disciplinary infractions for fighting/threatening other inmates/staff	Yes/No

Table A5. History of Noncompliance Subscale[17]

Subscale Items	Values
On probation or parole at time of current offense[18]	**Probation/Parole/Both/Neither[19]**
Number of parole violations	0, 1, 2, 3, 4, 5+
Number of times person has been returned to prison while on parole	0, 1, 2, 3, 4, 5+
Number of new charge/arrests while on probation	**0, 1, 2, 3, 4, 5+**

Number of probation violations/revocations	**0, 1, 2, 3, 4, 5+**

Table A6. Criminal Involvement Subscale[20]

Subscale Items	**Values**
Number times offender has been arrested as adult/juvenile for criminal offense	**Any value accepted**
Number times offender sentenced to jail for ≥30 days	**0, 1, 2, 3, 4, 5+**
Number of new commitments to state/federal prison (include current)	**0, 1, 2, 3, 4, 5+**
Number of times person sentenced to probation as adult	**0, 1, 2, 3, 4, 5+**

Table A7. Vocation/Education Subscale[21]

Subscale Items	**Values**
Completed high school diploma/GED	Y/N
Final grade completed in school	—
Usual grades in high school	A, B, C, D, E/F, did not attend
Suspended/expelled from school	Y/N
Failed/repeated a grade level	Y/N
Currently have a job	Y/N
Have a skill/trade/profession in which you can find work	Y/N
Can verify employer/school (if attending)	Y/N

Amount of time worked or in school over past 12 months	12 months full time, 12 months part time, 6+ months FT, 0-6 months PT/FT
Feel that you need more training in new job or career skill	Y/N
If you were to get (or have) a good job, how would you rate your chance of being successful?	Good, Fair, Poor
How hard is it for you to find a job ABOVE minimum wage compared to others?	Easier, Same, Harder, Much Harder

Table A8. Substance Abuse Subscale[22]

Subscale Items	**Values**
Do you think your current/past legal problems are partly because of alcohol or drugs?	Y/N
Were you using alcohol when arrested for your current offense?	Y/N
Were you using drugs when arrested for your current offense?	Y/N
Are you currently in formal treatment for alcohol/drugs?	Y/N
Have you ever been in formal treatment for alcohol/drugs?	Y/N
Do you think you would benefit from getting treatment for alcohol?	Y/N
Do you think you would benefit from getting treatment for drugs?	Y/N
Did you use heroin, cocaine, crack, or methamphetamines as a juvenile?	Y/N

References

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 35–44, New York, NY, USA. Association for Computing Machinery. 10.1145/3097983.3098047.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer,M., and Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(234):1–78.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. Technical report, ProPublica.

Baradaran, S. (2013). Race, prediction, and discretion. *The George Washington Law Review*, 81:157.

Barnes, G. C. and Hyatt, J. M. (2012). Classifying adult probationers by forecasting future offending. Technical Report 238082, National Institute of Justice, U.S. Department of Justice.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*. 10.1177/0049124118782533.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40. 10.1177/0093854808326545.

Bushway, S. D., Owens, E. G., and Piehl, A. M. (2012). Sentencing guidelines and judicial discretion: Quasi-experimental evidence from human calculation errors. *Journal of Empirical Legal Studies*, 9:291–319. 10.1111/j.1740-1461.2012.01254.x.

Bushway, S. D. and Piehl, A. M. (2007). The inextricable link between age and criminal history in sentencing. *Crime & Delinquency*, 53(1):156––183. 10.1177/0011128706294444.

Citron, D. (2016). (Un)fairness of risk scores in criminal sentencing. *Forbes, Tech section*.

Coglianese, C. and Lehr, D. (2018). Transparency and algorithmic governance. Technical report,Working Paper. Penn Program on Regulation, University of Pennsylvania Law School.

Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post (op-ed)*.

Crow, M. S. (2008). The complexities of prior record, race, ethnicity, and policy: Interactive effects in sentencing. *Criminal Justice Review*, 33(4):502–523. 10.1177/0734016808320709.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity: Performance of the COMPAS risk scales in Broward County.

Fairness, Accountability, and Transparency in Machine Learning (FATML) (2018). https://www.fatml.org. [Online; accessed 10-June-2018].

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Flores, A. W., Lowenkamp, C. T., and Bechtel, K. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals". *Federal Probation*, 80(2).

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Gottfredson, S. D. and Jarjoura, G. R. (1996). Race, gender, and guidelines-based decision making. *Journal of Research in Crime and Delinquency*, 33(1):49–69. 10.1177/0022427896033001004.

Helmus, L., Thornton, D., Hanson, R. K., and Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research & Treatment*, 24(1):64–101. 10.1177/1079063211409951.

Ho, V. (2017). Miscalculated score said to be behind release of alleged twin peaks killer. *SFGate (San Francisco Chronicle)*.

Hoffman, P. B. and Adelberg, S. (1980). The Salient Factor Score: A nontechnical overview. *Federal Probation*, 44:44.

Howard, P., Francis, B., Soothill, K., and Humphreys, L. (2009). OGRS 3: The revised offender group reconviction scale. Technical report, Ministry of Justice.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., and Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice and Behavior*, 34(1):37–59. 10.1177/0093854808326545.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Technical report, ProPublica.

Lowenkamp, C. T. and Latessa, E. J. (2004). Understanding the risk principle: How and why correctional interventions can harm low-risk offenders. *Topics in Community Corrections*, 2004:3–8.

Nafekh, M. and Motiuk, L. L. (2002). The statistical information on recidivism, revised 1 (sirr1) scale: A psychometric examination. Technical Report PS83-3/126E-PDF, Correctional Service of Canada. Research Branch.

Netter, B. (2007). Using group statistics to sentence individual criminals: an ethical and statistical critique of the Virginia risk assessment program. *The Journal of Criminal Law and Criminology*, pages 699–729.

Northpointe (2012, 2015, 2019). Practitioner's Guide to COMPAS Core. Technical report, Northpointe, Inc.

Northpointe Inc. (2009). Measurement & treatment implications of COMPAS core scales. Technical report, Northpointe Inc. https://www.michigan.gov/documents/corrections/Timothy Brenne Ph.D. Meaning and Treatment Implications of COMPA Core Scales 297495 7.pdf. Accessed: February 2, 2020.

Patrick, K. (2018). Should an algorithm determine whether a criminal gets bail? *Inside Sources*.

Pennsylvania Commission on Sentencing (2018). Risk Assessment Project Phase III: Racial Impact Analysis of the Proposed Risk Assessment Scales. Technical report, Pennsylvania.

Petersilia, J. and Turner, S. (1987). Guideline-based justice: Prediction and racial minorities. *Crime & Justice*, 9:151. 10.1086/449134.

Redding, R. E. (2009). Evidence-based sentencing: The science of sentencing policy and practice. *Chapman Journal of Criminal Justice*, 1:1–19.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215. 10.1038/s42256-019-0048-x.

Rudin, C. and Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review,* 1(2). 10.1162/99608f92.5a8a3a3d.

Rudin, C. and Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *INFORMS Journal on Applied Analytics*, 48(5):449–466. 10.1287/inte.2018.0957.

Stevenson, M. T. and Slobogin, C. (2018). Algorithmic risk assessments and the double-edged sword of youth. *Washington University Law Review*, 96. 10.2139/ssrn.3225350.

Tan, S., Caruana, R., Hooker, G., and Lou, Y. (2018). Distill-and-compare: Auditing blackbox models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 303–310, New York, NY, USA. Association for Computing Machinery. 10.1145/3278721.3278725.

Tollenaar, N. and van der Heijden, P. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):565–584. 10.1111/j.1467-985X.2012.01056.x.

Turner, S., Hess, J., and Jannetta, J. (2009). Development of the California Static Risk Assessment Instrument (CSRA). University of California, Irvine, Center for Evidence-Based Corrections.

Ustun, B. and Rudin, C. (2019). Learning optimized risk *scores. Journal of Machine Learning Research*, 20(150):1–75.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Wahi, M. M., Parks, D. V., Skeate, R. C., and Goldin, S. B. (2008). Reducing errors from the electronic transcription of data collected on paper forms: A research data case study. *Journal of the American Medical Informatics Association*, 15(3):386––389. 10.1197/jamia.M2381.

Westervelt, E. (2017). Did a bail reform algorithm contribute to this San Francisco man's murder? NPR, August 8.

Wexler, R. (2017a). Code of silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out. *Washington Monthly*.

Wexler, R. (2017b). When a computer program keeps you in jail: How computers are harming criminal justice. *New York Times*.

Zeng, J., Ustun, B., and Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722. 10.1111/rssa.12227.

Footnotes

1. The capitalization is directly from Northpointe Inc. (2009). ↩
2. https://github.com/propublica/compas-analysis ↩
3. COMPAS subcales are inputs to the recidivism scores. We have some but not all of the questionnaire features that determine each subscale. For one of the History of Noncompliance features, "Was this person on probation or parole at the time of the current offense?," we can determine only whether the person was on probation. ↩
4. In all but one case, the subscale inputs are binary or count variables that take up to only 6 values. The exception is the Criminal Involvement Subscale, which takes the total number of prior arrests, and for which we have data to directly validate. ↩
5. We are trying to determine whether the COMPAS remainder (general COMPAS after subtracting the main age terms) still depends on age. The numbers for with age look very similar to the numbers without age. Thus, age does not seem to participate in the remainder term because accuracy does not change between the two rows. ↩
6. Again, age does not seem to participate in this remainder. ↩
7. There is little difference with and without race. The differences between algorithms are due to differences in model form. Age at COMPAS screening date and age-at-first-arrest are included as features. ↩
8. Age at COMPAS screening date and age-at-first-arrest are included as features. ↩
9. Number of charges (resp. # Arrests) counts the prior charges (resp. arrests) up to but not including the current offense (i.e., the offense we believe triggered the COMPAS score calculation), since this is how COMPAS counts prior offenses. However, the current offense may be included in Select Prior Charges. Note that any Subsequent Crimes beyond the 2-year mark of the COMPAS score calculation or outside of Broward County may not be contained in our database. Note that if an individual spent time in prison after the COMPAS calculation date, subsequent charges they generated while in prison may not be in our database. The notation "(F/M, N)" next to each charge

gives the charge degree (F = Felony or M = Misdemeanor) and the number of instances in this charge (N). ↩

10. Figure A1 in the Appendix plots the probability of 2-year recidivism (defined by arrest within 2 years) as a function of age for individuals in Broward County, Florida, showing how it decreases as a function of age. ↩

11. See the Appendix for full distributions. ↩

12. Age at COMPAS screening date and age-at-first-arrest are included as features. Unlike when predicting the COMPAS raw score remainder, we include the current offense in criminal history features. ↩

13. Age at COMPAS screening date and age-at-first-arrest are included as features. Unlike when predicting the COMPAS raw score remainder, we include the current offense in criminal history features. ↩

14. The significance levels are not valid, since the model assumptions with respect to age are broken. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. Our results are based on 5759 observations while ProPublica's results are based on 6172 observations. ↩

15. https://github.com/beauCoker/age_of_unfairness ↩

16. We compute the components in bold font. We do not have the data to compute the other components. The feature for family violent arrests is always 0 so it is not useful for prediction. We classify a charge as family violence if the statute is 741.28, which corresponds to the definition of domestic violence in the Florida crime code. In our data set there were no instances of this statute. ↩

17. We compute the components in bold font. We do not have the data to compute the other components. ↩

18. Only "On Probation" and "Not On Probation" computed. ↩

19. See Note 17. ↩

20. We computed all the components. To compute the number of arrests component, we interpreted a charge as an arrest. ↩

21. We do not have the data to compute any of these components. ↩

22. We do not have the data to compute any of these components. ↩