Harvard Data Science Review

Broader Issues Surrounding Model Transparency in Criminal Justice Risk Scoring

Cynthia Rudin[1], Caroline Wang, Beau Coker

[1]Professor of Computer Science, ECE, and Statistics, Duke University

Published on: Mar 31, 2020

This is a rejoinder to the invited commentaries on the article, "The Age of Secrecy and Unfairness in Recidivism Prediction."

We are pleased to have the opportunity of this rejoinder, which permits a broader discussion on the issues surrounding proprietary models than in our article, "The *age* of secrecy and unfairness in recidivism prediction." We would like to thank all discussants, and the Editor for organizing the discussion. We will start our discussion with the response to Eugenie Jackson and Christina Mendoza, who represent Northpointe, the company that owns COMPAS.

1. Eugenie Jackson and Christina Mendoza from Northpointe/Equivant: On Proprietary Models and Reproducibility

We would like to thank Northpointe/Equivant for engaging with us. We appreciate that it may have been difficult to do so under these circumstances.

In our manuscript, we provided evidence of a possible error in the COMPAS Practitioner's Guide. The COMPAS Practitioner's Guide indicates a linear dependence on age, when data from Florida indicates that the dependence is nonlinear. We illustrated that the lack of transparency of models like COMPAS can have serious consequences, for instance, leading to incorrect downstream analyses, such as ProPublica's Machine Bias article.

As it seems from Northpointe's response, *age may indeed have been transformed nonlinearly, in agreement with our paper's conjectures, and contradicting the COMPAS Practitioner's Guide.*

Specifically, Northpointe states: "Among the deficiencies in the authors' arguments [...] The authors have taken a clearly informal description of the VRRS score in the Practitioner's Guide to COMPAS Core (Northpointe Inc., 2019) for a complete technical description of the VRRS model. This guide is written for practitioners and is not intended to be a technical document. Discussions of appropriate variable transformations are beyond its scope and would not further its goals..." We contrast this with what is actually written in the Practitioner's Guide:

> $$\begin{aligned}&\text{Violent Recidivism Risk Score}\\&= (\text{age}*-\text{w})+(\text{age-at-first-arrest}*-\text{w})+(\text{history of violence} * \text{w})\\&\quad+\quad(\text{vocation education} * \text{w}) + (\text{history of noncompliance} * \text{w}) \quad (1)\end{aligned}$$

Is this truly an informal description? It is an equation! Might the correct version of this equation be:

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

where *f* and *g* are proprietary transformations of age, such as linear splines? Is interpreting Equation (1) as the precise calculation used within COMPAS really a ‘deficiency in our argument’? Might practitioners (or reporters) make the same “error” (Jackson and Mendoza, 2020) we did, if this is written in the Practitioner's Guide? We believe this is definitely a possibility.

The major questions we consider in our article, not sufficiently addressed by Northpointe's response, still remain:

Do we need COMPAS, or other proprietary models for high stakes decisions about individuals? COMPAS is a black box *because* it is proprietary. We do not know what computation it is performing. If COMPAS is indeed simply logistic regression with a small set of transformed features, as Northpointe claimed in its response, then their legal team has gone to quite a lot of trouble to protect simple logistic regression, battling in court to protect proprietary secrets at the expense of individuals whose score might have been miscomputed. Many have expressed concerns about due process rights, as argued in Loomis vs. Wisconsin (2016).

If COMPAS is just logistic regression with a few transformed variables, do we really need a company to create such a model? Why not use the other numerous recidivism prediction tools for pretrial release and parole decisions that are not proprietary? Why not have academics perform this work if it is logistic regression on a few well-defined variables? Perhaps the transformation of COMPAS' variables is somehow special. But in that case, why doesn't COMPAS perform better than other models for its risk assessments? Indeed, we know that even the most powerful and complicated (black-box) machine learning tools don't seem to perform any better than even very simple rules for recidivism prediction (Angelino et al. 2018, Zeng, Ustun, & Rudin, 2017) or handcrafted logistic regression models (Tollenaar and van der Heijden, 2013, 2019). So why, then, do we need COMPAS, or models like it, for risk assessment? Our point extends beyond criminal justice to other domains where companies provide algorithmic tools for high stakes decisions that deeply affect lives. If a company is selling a product for which (1) black boxes provide no clear empirical benefit and (2) there can be serious consequences for not understanding these models in practice (such as typos influencing decisions, due process concerns, health consequences, or confusion over whether they are racially biased), then perhaps we should not be using such models.

Northpointe's argument (as we understood it) is that there are many different scores in the COMPAS suite, each meant for a different purpose, and each based in criminological theory. Theory, however, must be supported by data (if the data exist). Here, the data do not support the models created with criminological theory any more than they support models created without theory, or models created with entirely different theories. There is no credible empirical evidence to show that this theory is serving its intended goal in practice, any more than the theory backing all of the publicly available risk scores (referred to by Desmarais in her response), or the machine-learning based risk scores.

The current theory embedded into COMPAS could also be wrong. In many regions of the U.S., age-crime curves have changed dramatically. If the theory underlying COMPAS reflects a strictly decreasing risk as a function of age (as we hypothesized in the article), then it may no longer be correct in many parts of the U.S. If the model remains proprietary, there could be serious ramifications for bail, parole, and sentencing across the country. A non-proprietary model would be easier to check for such underlying problems, and thus it would be easier to repair.

Some of the evidence cited by COMPAS to support it are academic papers that are not reproducible because the formulas and/or data are not publicly available. Even Northpointe in their response state that "care and due diligence are critical ingredients to the procedure of academic research and publishing" (Jackson and Mendoza, 2020). But without anyone else being able to verify the analysis, how can such care be made? In contrast, all of our analysis is reproducible. So is ProPublica's. In fact, the only reason we were able to figure out the flaws in the ProPublica analysis is because they made the data and the code public.

Northpointe states, "A feature that has been widely ignored is that every agency using these instruments already has full access to all risk variables, logic, scoring processes, and guidelines for the appropriate uses of risk assessment procedures" (Jackson and Mendoza, 2020). Here, "the agency" does not include the person whose score is being computed. In the legal briefs of the Loomis vs. Wisconsin case, within each person's Presentence Investigation Report (PSI), they receive only a statement that the COMPAS calculation is proprietary, rather than any information on how their score is computed: "All PSIs containing a COMPAS risk assessment must include 'a written advisement listing [its] limitations' and 'should inform sentencing courts' that (1) '[t]he proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are determined' " (Loomis vs. Wisconsin, 2016).

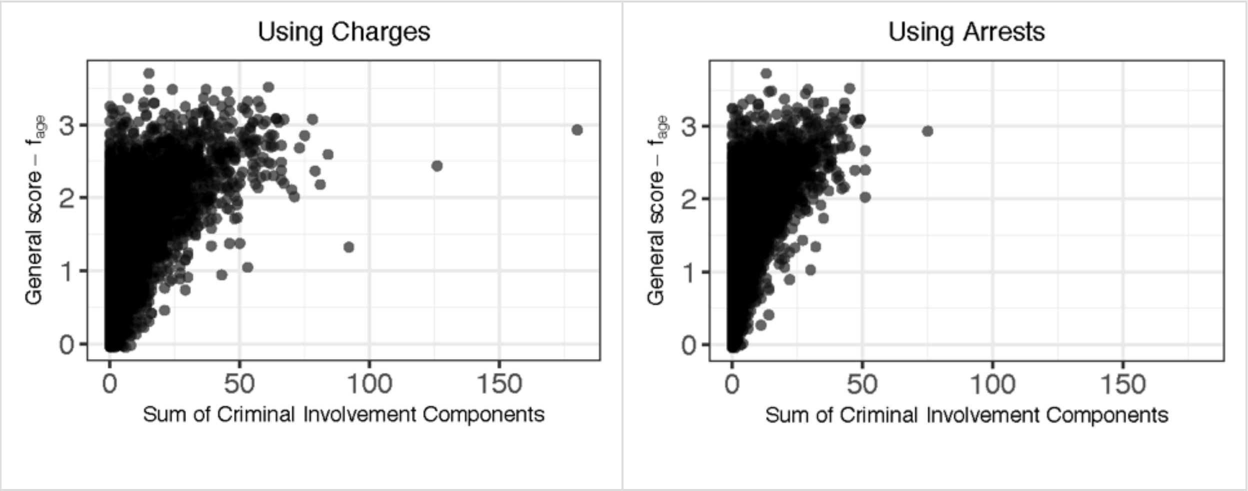
Worse, there are more pressing practical issues—in practice, sometimes people are assessed using the wrong COMPAS score (e.g., pretrial release decisions are made using the parole score in Broward County), or perhaps data errors prevent them from receiving the correct prediction. As discussed earlier, there have been concerns raised about due process with these proprietary scores.

We were criticized for stating that COMPAS is based on 137 variables, and for questioning whether COMPAS truly requires all of the variables it collects. Indeed, the VRRS only depends on up to 26 of the 137 variables, as shown in Table 1 in our paper. However, the accuracy of the COMPAS violent and recidivism risk scores seems to be approximately equivalent to that of models that use far fewer features. If these models in the COMPAS suite do not require all of their collected variables to predict accurately, perhaps the other COMPAS scales do not require all their variables either. Given that some of these 137 variables contain highly sensitive, private information (such as whether one's mother was arrested), perhaps they should not be collected if they are not absolutely essential.

Northpointe indicated that there are variables correlated with age that remained in the data used to train the machine learning models for predicting the COMPAS scores minus age splines. Specifically they say, "In Section 2.2 they compare the performance of several models both with and without age inputs, but always with the 'age at first arrest' input" (Jackson and Mendoza, 2020). We investigated age-at-first-arrest, and were not able to extract any dependence on this variable. It is possible that there is a complicated relationship between age-at-first-arrest and criminal history, however, according to the COMPAS Practitioner's Guide, there are no subscales that combine age-at-first-arrest and criminal history, and we were not able to find a dependence on age-at-first-arrest alone.

Northpointe also stated that the features in COMPAS are computed using arrests, rather than charges. In reworking our analysis with arrests[1], we potentially came closer to understanding COMPAS' dependence on criminal history, as shown by the clearer lower bound in Figure 1 (right), though cleaner criminal history data may make this dependence clearer. To carry this analysis further, if given cleaner criminal history data, one might be able to construct an approximation of the Criminal Involvement Subscale as the lower bound of Figure 1 (right). Based on this figure, we believe it is possible that with cleaner criminal history data, it may be possible to recover COMPAS' dependence on criminal history. At that point in the analysis, it may not be possible to proceed further without data to calculate the other two COMPAS general score subscales: Vocation/Education and Substance Abuse.

COMPAS general remainder, calculated using charges (left figures) and arrests (right figures):



Here are the same figures, zoomed in:

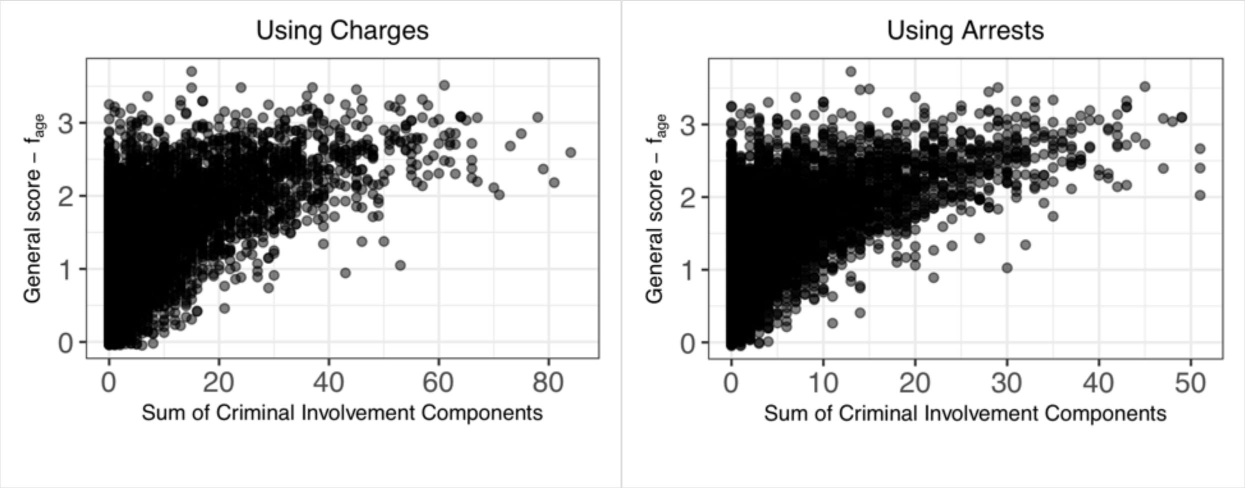


Figure 1: The COMPAS general remainder (COMPAS general score - f_{age}) plotted against the unweighted sum of terms in the Criminal Involvement Subscale (CIS). Left plots: The unweighted sum of criminal involvement components is computed using prior charges. Right plots: The sum is instead computed using prior arrests. Age outliers (from the analysis in the article) are omitted. Points are semi-transparent. Considering the lowest points for each CIS score, a somewhat clearer lower bound emerges in the right subfigures than in the left subfigures. High values of CIS, towards the right in all of the plots, may yield noisier data because these individuals have many past crimes, leading to error-prone values.

Northpointe argues that we “conflate complexity of method, complexity of model, interpretability, and secrecy,” and suggests that a branch and bound algorithm is not simple or interpretable (Jackson and Mendoza, 2020). In machine learning terminology, an algorithm produces a model, and the model makes predictions. It is only the complexity of the *model* that matters, not the algorithm producing it. Let us say that the model is: ‘if (age=19-20 and sex=male) OR (age=21-22 and priors=2-3) OR (priors >3) then predict arrest, else predict no arrest.’ Regardless of what algorithm came up with that model (whether it is the branch and bound scheme of CORELS, or the gradient descent method of logistic regression, or a brute force search over model space), it is a very small interpretable model. While we are not recommending using the model above (which indeed came from the CORELS algorithm), the fact that we can see what the model is computing allows us to critique it, rather than to trust it blindly.

For us, a model is a secret if it is a function we do not know. It seems that we and Northpointe stood on opposite sides of this debate, but we appear to be coming closer, as they are trying to make their model public under patent. This is a big change!

2. Shawn Bushway: How Interpretable Models Interact with Human Decision Makers

Our compliments to Bushway for his elegantly-written response and useful historical context to risk scores in criminal justice.

We agree with Bushway's critiques of risk scores. To clarify, *for no application (pretrial release, bail, sentencing, parole, etc.) do we advocate for decisions that depend only on age and criminal history.* We advocate only for interpretability and transparency. As we noted, one of the main reasons for risk scores to be transparent is to allow decision makers to determine how to include (and calibrate) extra factors that are not in a database. Due to the lack of data, part of the decision actually does need to be made by a black box (namely, a human). The question then becomes how much freedom can or should be given to adjust the risk scores. For instance, if only additional *mitigating* factors can be included, the risk estimates can only go down from the baseline of a statistical prediction, never up, which is an important way that risk scores and human decision makers can work together within their limitations. Alternatively, if factors outside of the database are limited in the way they can change risk scores, then human decision makers cannot heavily influence the decision.

Bushway's comments on the problems caused by the cycle of crime are well-taken. Indeed, if decisions depend entirely on criminal history, then we do not permit potentially important dynamic factors (such as participation in a reentry program) to influence decisions. However, *lack of reliable data* is problematic. If all individuals in the database were involved in the same job reentry program, for example, and this program records data in a reliable way, then we should use that data. Doing so might alleviate Bushway's concern for the inclusion of dynamic factors. The problem is that in many places (perhaps in most places), either the data do not exist or the data are not reliable. Incorporating known (but not recorded) dynamic factors is where judicial discretion can be helpful in adjusting the estimated risk. However, it is important that the judges have a useful baseline to start. No judge can calculate risks from a large database in their head. With the anchor of a risk score based on data and statistics, and with these statistical models controlled for fairness considerations, a decision maker can create a more informed decision, *and any modification from these statistical models can be quantified,* even if the modification is performed manually by a human.

Some have recently argued (Barabas et al., 2019; Barabas, Dinakar, & Doyle, 2019) that even risk assessments for pretrial release decisions should not be used at all, mainly for reasons about fairness and quality of data; however, we disagree, based on the arguments in the rebuttal by Desmarais et al. (2019). The strongest argument (mentioned also by Bushway) is that the alternative—judges deciding without data—leads to black-box decisions that are even more biased, inconsistent, and unfair.

Statistics can be adjusted to correct for problems with fairness *uniformly*, whereas judges are inconsistent and not so adjustable.

Bushway comments that models wane in effectiveness over time as new treatment programs come into effect. A particularly useful aspect of interpretable machine learning tools is that they allow faster creation of interpretable risk scores directly from data, as well as faster critique and development of these models. Thus, as new treatment programs alter the landscape of data generation, machine learning models can be easily retrained to change along with it.

Interpretable machine learning tools change the nature of the discussion of constructing risk prediction tools. Since machine learning tools create an optimal solution for a specific balance of our values (such as accuracy, fairness, etc.), these tools encourage us to argue about the *balance of our values*, rather than the more technical questions such as what coefficients we should choose for our models.

3. Brandon Garrett: There is No Winner between a Black-Box Algorithm and a Black-Box Human

Brandon Garrett's expertly-written response touches on many different and important points.

Garrett draws attention to an irony in the Wisconsin Supreme Court decision involving COMPAS. He writes: "In its *Loomis* decision in 2016, the Wisconsin Supreme Court rejected the claims, emphasizing that the risk assessment information was only advisory, and judges have discretion to consider or discount the recommendation provided by the instrument. Yet, with the algorithm treated as a trade secret, the judge does not understand how the risk score is calculated any better than the defense" (Garrett, 2020).

There are several layers of interesting reasoning in that concise statement. First, if the judge does not know how much age or criminal history contributes to a risk calculation, how can they be expected to gauge outside factors on the same scale? For instance, if the judge does not know how many points the defendant receives for age or criminal history, can the judge then incorporate information that is not in the database, such as the person's participation in a new drug treatment program, or other involvement in the community? If the database's factors and outside factors are not shown on the same scale, it can be challenging for the judge to adjust the risk calculation in an effective way. This can make the resultant combination of the judge and the risk calculation untrustworthy. If the judge does not attempt to combine outside information with the black-box risk score, then there are two remaining options: either (1) the judge does not use the risk calculation, in which case the whole decision then returns to the black box of judicial discretion; or (2) the judge uses (or slightly adjusts) the black-box risk calculation, so that the decision is made mainly by the black-box risk calculation, which neither the judge nor the defendant understands. In that case, the black-box risk calculation, which is supposed to play only an advisory role, becomes the decision maker. In both of these cases, the fact that the model is proprietary forces the use of a black box in the decision.

Garrett makes an extremely important point regarding openness of data. In our studies (Rudin et al., 2020; Zeng, Ustun, & Rudin, 2017), we have used only publicly available data. We echo Garrett's call to action on open science and open data. We are hopeful that our paper might serve as a loudspeaker on this topic, rather than just an echo.

4. Sarah Desmarais: Is Our Point Obvious? Not to Everyone.

Sarah Desmarais' clear, succinct, and carefully-crafted perspective is much appreciated. We agree with all points.

On using arrests (or charges) instead of convictions: In our current study, we were replicating COMPAS and ProPublica, so we were using the features they used. We agree that conviction counts are a better measure for risk scores. We also do not have complete conviction data.

One interesting point brought up by Desmarais is that COMPAS is getting too much attention: it has obvious flaws, and there are numerous alternatives that are not as flawed. In some ways, it is obvious to experts that we do not need COMPAS. Then why is this paper needed? Wasn't its main point—that transparency and interpretability is paramount—obvious in the first place?

Actually, our main point may be obvious to experts on criminal justice, but may not be obvious to others. As we write this, there is a large and growing algorithmic fairness literature. Much of this literature grew essentially from the ProPublica Machine Bias article. Almost none of this literature considers model transparency or interpretability. There is also an ever-increasing number of researchers who aim to explain black boxes, rather than replace them with transparent and interpretable models. There is a chasm between an interpretable model and an 'explained' black box (see Rudin, 2019; Rudin & Radin, 2019). As Brandon Garrett (2020) writes, "The lesson that simple and interpretable risk assessments can be used, rather than an error-prone, complex, and inscrutable risk system, has been lost on many criminal justice decision makers."

5. Greg Ridgeway: Data Privacy Concerns are Real.

Greg Ridgeway has added some critical, important details and clarity around some of our main points. Indeed, transparency *does* reduce the risk of misunderstanding.

Ridgeway writes that we misunderstand COMPAS' purpose as only a risk assessment—neglecting its nature as a needs assessment. Actually, we do recognize it as having a needs assessment component; we questioned whether the comprehensive additional information being collected is actually *useful* for the needs assessment.

Are each of the 130+ survey questions needed to determine the needs of individuals? Some survey questions may have an adverse effect on the privacy of individuals. Survey questions for the needs assessments can be much broader than those of the risk assessment, and their answers can contain information about individuals who are *not* being assessed (e.g., COMPAS' Core Risk Assessment from Wisconsin requires the criminal history of individuals' family members [Sample, 2011]). Are we certain, for instance, that in order to assess risks or needs, it is necessary to collect data on exactly which immediate family members have committed a crime in the past or have a drug addiction? We must be sure that the value of collecting this information offsets the privacy concerns of collecting it.

We consider whether there could be alternative survey questions that are equally effective but raise fewer privacy concerns. This might involve less direct questions, such as some of those used in the COMPAS Substance Abuse Subscale: "Do you think you might benefit from getting treatment for drugs?" or questions like those in the Vocation/Education Subscale "[Do you] feel that you need more training in new job or career skill?" Questions like this do not probe for private information about specific individuals.

Beyond mitigating privacy concerns, smaller surveys might be easier to troubleshoot. If these surveys could be simplified by reducing the overall number of questions, it could lead to reduced costs (because smaller surveys are less costly to conduct), and reduced typographical errors (because smaller surveys lead to fewer chances for errors), and thus a more efficient overall system. Perhaps even only one or two well-chosen questions would be sufficient to determine relevance for a particular treatment program, but we will not know that until we attempt to identify such questions.

6. Alexandra Chouldechova: On Applying Classical Decision Analysis to Criminal Justice Decisions

Chouldechova redirects the reader back to fairness considerations of the type that ProPublica was concerned with in writing their article: notions of bias against specific groups. She argues that transparency might require a trade-off with other forms of fairness, and also argues that the type of transparency we consider (simplicity in model) is not sufficient to address certain types of bias. While we agree with the major principles that Chouldechova advocates for (fairness is important, overall transparency is important beyond predictions), let us dive into some of Chouldechova's examples, where we can clarify possible differences.

On sentencing: Our work is primarily focused on pretrial release and parole decisions, not sentencing decisions. Our data are not from sentencing decisions. Risk prediction indeed plays a minor role in sentencing. In the movie *Minority Report*, individuals are punished based on the crimes they might commit in the future, not the crime they have committed in the past, but in real criminal sentencing in America, punishment is primarily based on crimes committed in the past; this is in accordance with the Model Penal Code (2017).

On decision making: Chouldechova points out that decision making cannot be based fully on automated risk predictions, which of course we agree with, hence the need for interpretability.

In the field of decision analysis, decision making is generally based on minimizing (possibly abstract) estimated *costs*, and incorporating future risks by weighing them according to the cost of each possible outcome. For instance, Lakkaraju and Rudin (2017) considered how one might design a tool that trades off among various costs involved in pretrial release decisions: the cost to society of having a released individual commit a crime before their trial, or not return for their trial; the cost of gathering information about individuals; the cost to taxpayers of keeping someone confined before their trial; and so on.

It would be useful to create a cost-benefit analysis per decision (data permitting). The costs that might be considered for each decision include costs for violations of fairness objectives, costs to society of future crime, costs to society of treatment programs, costs to families involved in the justice system, and costs for not providing sufficient punishment for the severity of the crime. The idea of a cost-benefit analysis leads to an important direction for future work: *placing modern criminal justice decisions in the framework of classical decision analysis*. Assuming all of the different aspects of each decision can be quantified, what combination of expected costs should we use to form an objective for decision making? Knowing this would bring us a big step closer to more consistent and informed decision making. However, enumerating the costs, let alone the risks, would be a major challenge.

On sacrificing simplicity for fairness: As Chouldechova points out, a lack of simplicity can manifest in several ways. However, in our view, the examples in this section of her response do not necessarily illustrate this point. As Chouldechova mentions, Breiman introduced the Rashomon effect, which is the observation that there are often a multitude of approximately-equally-accurate models for a given dataset. Given a large set of equally-accurate models, it is possible that one or more of them can exhibit additional desirable properties, such as simplicity or fairness (this technical argument is spelled out by Semenova, Rudin, & Parr, 2019). Because of the Rashomon effect, it may be possible to attain a desired balance among accuracy, simplicity and/or fairness. Kleinberg and Mullainathan (2019) overlook the Rashomon set entirely by asserting that there is a complex (possibly black-box) function *f* that we must approximate. This *f* is a model—an “admissions rule”—that may have been previously trained on the data. In other words, the authors have already selected a model from the Rashomon set, and now aim to approximate it. Rather than attempting to approximate the complicated model *f*, the practitioner may have been able to find an alternative model from the Rashomon set in the first place, which achieved the desired balance of accuracy, simplicity, and/or fairness. Assuming that a complex model *f* is the best solution for the problem at hand, and that it is necessary to approximate it—rather than trying to find a simpler, fairer model initially—is a common, but problematic approach (see, e.g., Rudin & Radin, 2019).

Chouldechova states that bias stemming from non-representative data “may be no easier to detect with a simple model than a black-box one,” but in many real cases, it *would* be much easier to detect bias with an interpretable model. For instance, would ProPublica have written their article if it was known publicly that COMPAS was a simple function of age and criminal history (which are unavoidably related to race)? Why didn’t they instead write their article about the numerous other risk scores used in the justice system whose formulas are not proprietary? We argue that it is precisely because COMPAS is a black box—and could potentially contain factors related to race other than age and criminal history—that ProPublica chose it.

Chouldechova points out that many types of transparency other than model interpretability would be more useful for determining whether a model is fair, citing the health insurance (Obermeyer, Powers, Vogeli, & Mullainathan, 2019) and Gender Shades (Buolamwini & Gebru, 2018) studies as examples where very simple statistics were sufficient to reveal obvious unfairness. However, for many applications, model interpretability is desirable for reasons other than determining traditional notions of fairness. As discussed in our responses to other discussants, model interpretability helps avoid consequences such as typos when computing risk scores, due process concerns, and helps judges use model predictions in conjunction with outside information.

7. Closing Thoughts: Lessons for Responsible Reporting

We would like to finish this rejoinder by addressing how the larger discussion on this topic began.

The ProPublica Machine Bias series has been highly acclaimed, winning awards for best reporting in 2016 from the Scripps Howard Foundation, and becoming a finalist for a 2017 Pulitzer for explanatory reporting. Their article is heavily cited in academic literature on fairness. In our work, we have presented substantial evidence that important parts of their analysis and conclusions were questionable.

ProPublica likely targeted COMPAS for an investigation of racism because of the main thing that differentiates it from other risk scores: the fact that it is not *transparent*. They did not discuss the many recidivism risk scoring systems in existence that are not black boxes. They did not discuss the fact that race is not needed for models to have predictions on par with the best machine learning models (Zeng, Ustun, & Rudin, 2017).

ProPublica argues COMPAS depends *more on race than on criminal history*, because they paired white and black individuals with similar numbers of priors (or where the white person had more serious priors) but where the white person had a much lower COMPAS score than the black person. They concluded COMPAS depends on race, other than through age and criminal history. They did not discuss the possibility that these pairs can arise from errors in the data or in the data processing, which are pervasive in criminal justice data, or from confounding with unobserved variables. In particular, they did not address the possibility that there are many inputs to the COMPAS score other than criminal history and age (not present in the ProPublica data) that could benevolently explain the differences in COMPAS scores between the white and black individuals they examined. Through lack of these careful investigations, ProPublica's findings cannot be trusted. In particular, the types of analyses they present can lead to an exaggerated illusion of racism.

Let us more closely consider one of ProPublica's cherry-picked pairs. In one of the pairs of individuals, ProPublica states that Gregory Lugo (white) has a COMPAS score of 1 (low risk) but prior offenses of three DUIs and one battery. Lugo's photo is placed next to that of Mallory Williams' (black), who has a COMPAS score of 6 (medium risk) but prior offenses of only two misdemeanors. However, querying Gregory Lugo's records from the Broward Clerk's database revealed only two cases with DUI charges, both of which were on the same day, and thus possibly corresponding to the same arrest. The discrepancy between COMPAS score and criminal history data that ProPublica observed could easily have resulted from data processing issues or data quality issues.

The ProPublica article leads to some lessons for responsible data science reporting:

1. Cherry-picked examples are sometimes misleading. They are capable of creating a strong illusion to support a false hypothesis. In this case, the hypothesis was that COMPAS depends on race other than through criminal history and age. In a database where errors are common, data processing can easily have mistakes, and where there are many unmeasured confounding variables, one might use cherry-picked examples to convincingly support a myriad of possibly false, misleading conclusions.
2. It is easy to fall into the trap of explaining a black-box model incorrectly. Incorrect explanations can lead to incorrect conclusions about whether these models depend on an important variable (like race).
3. Reporters should communicate their findings to and consult domain experts prior to publishing. Criminologists may have been able to warn ProPublica that their results are questionable, but were not consulted, according to Flores et al. (2016).

The ProPublica article demonstrated the value of reproducibility in science. The data provided through ProPublica, supplied by the Broward County Sheriff's office, along with their public code, has allowed the possibility of uncovering problems with their analysis.

By adding an extra level of analysis to COMPAS and the ProPublica study, we hoped to highlight what had actually happened here—that an avalanche of work on algorithmic fairness stemmed from a *lack of transparency*. Transparency is critical to fairness, error-checking, and understanding. Fairness is important, but we cannot forget that it relies upon a foundation of transparency.

Discussion

Read invited commentary by:

- Shawn Bushway (The RAND Corporation)
- Alexandra Chouldechova (Carnegie Mellon University)
- Sarah Desmarais (North Carolina State University)
- Brandon L. Garrett (Duke University School of Law)
- Eugenie Jackson and Christina Mendoza (Northpointe, Inc.)
- Greg Ridgeway (University of Pennsylvania)

References

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Certifiably optimal rule lists for categorical data, *Journal of Machine Learning Research.* 18(234), 1—78. http://jmlr.csail.mit.edu/papers/v18/17-716.html

Barabas, C., et al. (2019). Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns. Unpublished. https://dam-prod.media.mit.edu/x/2019/07/16/TechnicalFlawsOfPretrial_ML%20site.pdf

Barabas, C., Dinakar, K., & Doyle, C. (2019). The Problems With Risk Assessment Tools. Opinion, *The New York Times.* https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, 77–91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

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

Desmarais, S., Garrett, B., & Rudin, C. (2019). Risk Assessment Tools Are Not A Failed 'Minority Report'. Perspectives, Law360. https://www.law360.com/articles/1180373

Flores, A.W., Lowenkamp, C. T., & 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. https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder

Kleinberg, J., & Mullainathan, S. (2019). Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability. 20th ACM Conference on Economics and Computation (EC). https://doi.org/10.3386/w25854

Lakkaraju, H., & Rudin, C. (2017). Learning Cost-Effective and Interpretable Treatment Regimes. Artificial Intelligence and Statistics (AISTATS). http://proceedings.mlr.press/v54/lakkaraju17a.html

Loomis v. Wisconsin, 881 N.W.2d 749 (Wis. 2016). https://www.scotusblog.com/wp-content/uploads/2017/02/16-6387-BIO.pdf

Model Penal Code, Sentencing 6B.09(3). (2017). American Law Institute.

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

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

The Pulitzer Prizes. (2017). Finalist: Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner and Terry Parris Jr. of ProPublica: For a rigorous examination that used data journalism and lucid writing to make tangible the abstract world of algorithms and how they shape our lives in realms as disparate as criminal justice, online shopping and social media. https://www.pulitzer.org/finalists/julia-angwin-jeff-larson-surya-mattu-lauren-kirchner-and-terry-parris-jr-propublica

Ridgeway, G. (2013). The pitfalls of prediction. *NIJ Journal*, 271, 34–40. https://nij.ojp.gov/topics/articles/pitfalls-prediction

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

Rudin, C, & 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). https://doi.org/10.1162/99608f92.5a8a3a3d

Rudin, C., Wang, C., & Coker, B. (2020). The *age* of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review,* 2(1). https://hdsr.mitpress.mit.edu/pub/7z10o269

Sample COMPAS Risk Assessment, Wisconsin. (2011). Submitted by Julia Angwin. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE

Scripps Howard Foundation. (2017). New Release: Scripps Howard Foundation recognizes excellence in journalism with 64th Scripps Howard Awards. March 7. https://scripps.com/wp-content/uploads/2019/04/SSP-2016-Scripps-Howard-Award-Winners-3-7-17.pdf

Semenova, L., Rudin, C., & Parr, R. (2019). A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv e-prints, 1908.01755. https://arxiv.org/abs/1908.01755

Tollenaar, N., & 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, 565. https://doi.org/10.1111/j.1467-985X.2012.01056.x

Tollenaar, N., & van der Heijden, P. (2019). Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PLoS One*, 14(3). ISSN 1932-6203. https://www.narcis.nl/publication/RecordID/oai%3Adspace.library.uu.nl%3A1874%2F379563

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

Footnotes

[1] Note that for one arrest, an individual might have multiple charges. Our raw data includes arrest data as well as charge data, but in our article we had used the charge information only, because the arrest data does not include the statute (the reason for the arrest), which is necessary to compute the Violence Subscale. Instead, in this rejoinder's analysis, we grouped all charges with the same date into one arrest record and used the statute data from the charges to characterize that arrest. We believe this is a reasonable assumption, as charges usually only occur when an individual is arrested, an individual is unlikely to be arrested more than once in a day (because arrested individuals are usually detained), and charge dates often reflect the date that the charges are filed (rather than the date that the charge is suspected to have occurred). This allowed us to use arrest rather than charge information to compute COMPAS subscale inputs. ↩