Investigating Human + Machine Complementarity for Recidivism Predictions

Sarah Tan
Cornell University

Julius Adebayo
MIT

Kori Inkpen
Microsoft Research

Ece Kamar
Microsoft Research

Abstract

When might human input help (or not) when assessing risk in fairness domains? Dressel and Farid (2018) asked Mechanical Turk workers to evaluate a subset of defendants in the ProPublica COMPAS data for risk of recidivism, and concluded that COMPAS predictions were no more accurate or fair than predictions made by humans. We delve deeper into this claim to explore differences in human and algorithmic decision making. We construct a Human Risk Score based on the predictions made by multiple Turk workers, characterize the features that determine agreement and disagreement between COMPAS and Human Scores, and construct hybrid Human+Machine models to predict recidivism. Our key finding is that on this data set, Human and COMPAS decision making differed, but not in ways that could be leveraged to significantly improve ground-truth prediction. We present the results of our analyses and suggestions for data collection best practices to leverage complementary strengths of human and machines in the fairness domain.

Introduction

The criminal justice field has used forecasting tools to perform risk assessment since the 1920s (see Gendreau, Freeze, and Goggin (1996) and Andrews, Bonta, and Wormith (2006) for meta-reviews). More recently, machine learning approaches are being used to inform bail and sentencing decisions (Berk and Bleich 2013; Angwin et al. 2016). But, this raises concerns around accuracy, fairness, and transparency of risk assessment systems (Andrews, Bonta, and Wormith 2006; Drake 2014; Zeng, Ustun, and Rudin 2016).

One ongoing debate is whether risk assessment systems are superior to human judgment. Grove et al. (2000) conducted a meta-analysis of 136 studies of human health and behavior to assess clinical (Human) versus mechanical (Machine) predictions. Their results revealed that on average, machine predictions were 10% more accurate than Human predictions, however there were some studies that showed no improvements and even a few cases where human predictions were more accurate. All three recidivism studies in Grove et al.'s analysis that tracked accuracy revealed similar levels of accuracy between human and machine predictions. On the other hand, Kleinberg et al. (2017) found that expert Humans (judges)'s decisions can sometimes be highly variable and biased by unobserved, irrelevant features.

In a recent study related to Human vs. Machine predictions of recidivism, Dressel and Farid (2018) showed that a widely used commercial risk assessment system for recidivism – COMPAS – was no more accurate or fair than predictions made by people with little to no experience in criminal justice. They sampled 1,000 defendants from the ProPublica COMPAS data (Larson et al. 2016) and asked Mechanical Turk workers to predict whether a defendant would recidivate within two years (the same label predicted by COMPAS). They also ran a second variant of their study where defendants' race was revealed. They did not find Human and COMPAS accuracies to be significantly different (COMPAS: 65.2%, Humans without defendant race information: 67.0%, and Humans with race information: 66.5%).

Although the Dressel and Farid study demonstrated that COMPAS and Human predictions were comparable, it was unclear whether COMPAS and Humans were accurate on the same or disjoint sets of defendants. Significant overlap would suggest that the Humans and COMPAS make similar assessments; less overlap suggests that human reasoning differed from machine analysis. Humans may have access to additional information or context not available to algorithmic systems; machines may not be influenced by the same biases that plague human judgment. Instead of focusing on the superiority (or lack thereof) of algorithmic systems compared to human judgment, we explore the similarities and differences between Human and COMPAS decisions to determine whether a hybrid approach that combines the strengths and addressess the weaknesses of human and machine decision making is possible.

Our contributions in this paper are:

- We analyze, on recidivism data, how human and machine decisions differ and how they make errors.
- We characterize areas of agreement and disagreement between human and machine decision making to better understand their complementarity.
- We investigate if hybrid models can leverage differences in human and machine decision making to improve recidivism prediction.
- Based on our findings, we discuss shortcomings of existing data sets and make recommendations for data collection best practices for future study of hybrid decision making in the fairness domain.

Table 1: Characterizing agreement and disagreement between COMPAS decisions, Human decisions, and ground truth. The number of defendants and characteristics for each of the eight cases are described.

Case	COMPAS Score	Human Score	Ground Truth	Agreement	Correctness	% Defendants	Feature Characteristics*
1	High	High	Yes	Agree	Both correct	49.0%	$1.5 <$ Priors ≤ 12.5
2	Low	Low	No	Agree	Both correct		$23.5 <$ Age ≤ 48.5 & Priors < 1.5
3	High	Low	Yes	Disagree	COMPAS correct	16.2%	$23.5 <$ Age ≤ 48.5 & Priors < 0.5
4	Low	High	No	Disagree	COMPAS correct		$1.5 <$ Priors ≤ 5.5 & Age >32
5	Low	High	Yes	Disagree	Human correct	15.9%	Similar to Case 4
6	High	Low	No	Disagree	Human correct		Similar to Case 3
7	High	High	No	Agree	Both incorrect	18.9%	No pattern, similar to Cases 1-6
8	Low	Low	Yes	Agree	Both incorrect		

* Characteristics determined by decision tree (Figure A3) and clustering analysis. See more details in Analysis and Results section.

Related Work

Humans and decision making. In addition to the work mentioned in the Introduction, Lakkaraju et. al., (2017b) showed that analyses of recidivism based on human decisions are further complicated by the "selective labels" problem, where observability of outcomes are affected by judges' release decisions. Other work studied how humans perceive different features as fair or not (Grgić-Hlača et al. 2018).

Hybrid models. Investigations across different domains identify that humans and machines have weaknesses and complementary abilities, thus suggesting benefits from hybrid models. In medicine, recent research showed that existing machine learning models with lower accuracy rates than human experts can decrease expert error rates by 85% (Wang et al. 2016). On challenging face recognition tasks, combining multiple expert opinions does not improve task accuracy, however complementing an expert with a inferior face recognition system can (Phillips et al. 2018). On the other hand, research on complementary computing demonstrated how humans and machines can be more effective together in problem solving (Horvitz and Paek 2007) and image classification tasks (Kamar, Hacker, and Horvitz 2012).

Diagnosing errors. The key to aggregating machine and human analyses to improve performance is understanding where and how machines and humans fail. Various approaches have been proposed for understanding where machine errors come from. Lakkaraju et al. (2017a) defined *unknown unknowns* as cases where the model is highly confident of its predictions but is wrong. Kulesza et al. (2015) uses human input to interactively correct a model. Another approach is to distill black-box model decisions to interpretable model classes to explain model failures (Nushi, Kamar, and Horvitz 2018; Tan et al. 2018). We follow a similar approach of utilizing interpretable models to analyze how machines and humans reason about recidivism, when and how their decisions differ and how they can be aggregated.

Approach

Constructing Human risk score

Our goal in this paper is to compare COMPAS and Human decision making. Before doing so, one may ask if the decisions made by Mechanical Turk workers in this data are *internally* consistent, or, in other words, if different Turk workers assess risk similarly for the same defendant. Large agreement among Turk workers increases confidence that our subsequent findings based on generating Human scores from Turk worker predictions generalize to Human decision making. We find that on average, 80% of the 20 Turk workers that assess the same defendant agree with each other. This is a high level of agreement, particularly for Mechanical Turk, where spam labeling is commonly observed (Ipeirotis, Provost, and Wang 2010). Hence, we perform a majority aggregation of Turk workers' predictions to assemble a Human risk score for recidivism risk, h_j. Specifically, we construct h_j by taking the mean prediction across 20 Turk workers for each defendant: let h_{ij} be Turk worker i's prediction for defendant j where $h_{ij} \in \{0, 1\}$, $i = 1, \ldots, 20, j = 1, \ldots, 1000$, we take $h_j = \sum_i h_{ij}/2$, dividing by two to scale h_j to 1-10, which is COMPAS' scale. We constructed scores for both conditions mentioned in the Introduction - a with-race Human score (HWR) for when Turk workers were told the defendants' race, and a no-race version (HNR).

For each score, we find the optimal cutoff point to binarize the score by computing calibration, false positive, and false negative rates at various cutoff points from 1 to 10. COMPAS, HNR, and HWR scores have approximately equal accuracy, false positive, and false negative rates at the cutoff point of $>= 5$ (Figure A5 in Appendix). Hence, we chose this cutoff point for all three scores. Note that Northpointe, the creator of COMPAS, also uses a $>= 5$ cut-off (Blomberg et al. 2010), and $>= 5$ is implied by Dressel and Farid's use of a "wisdom-of-the-crowd" based majority rules criterion.

Partitioning by agreement and correctness

We now sketch our approach towards studying how COMPAS and Human scores agree or disagree, and interact with ground truth. Table 1 describes eight possible combinations of two binary risk scores and ground truth. These eight combinations can be grouped into the four partitions illustrated in Table 1: Both correct, Both incorrect, Human correct, and COMPAS correct.

Comparing the level of agreement and correctness between the Human and COMPAS scores, we found that al-

most 50% of the time, Humans and COMPAS agree and are correct (Table 1). However, for the remaining 50% of defendants, either one, or both scores were incorrect. This suggests that if error regions of COMPAS and Humans do not perfectly overlap **and can be characterized**, then decision-making processes can potentially be improved through utilizing the complementary views of humans and machines.

When both risk scores agree and are correct, either score will return the same prediction, hence it does not matter which is used (in terms of accuracy). The space where both scores agree but are incorrect according to ground truth is a blind spot for COMPAS and Humans, also called *unknown unknowns* (Lakkaraju et al. 2017a). To characterize the space of agreement or disagreement between COMPAS and Human scores, we use clustering and decision trees. Table 1 summarizes our findings of the features that characterize each case. Finally, when COMPAS and Human scores disagree (Cases 3-6 in Table 1) we train hybrid risk scoring models to see if they can leverage disagreement between the two scores to improve on the accuracy of single scores.

Hybrid models

The simplest hybrid model is an average of two risk scores. We train a slightly more sophisticated model - a **weighted average** hybrid model that learns the optimal linear combination of two risk scores to predict ground truth. We also train **direct** and **indirect** hybrid models: direct models directly predict ground truth recidivism as a function of defendant features and Turk worker features (race, gender, and age) and the two scores; indirect models first predict which of the two scores to pick, then take that score's prediction of ground truth. We test the hybrid models against **random** and **single** score baselines. We use two types of random baselines: random ground truth labels, and random risk score. Single score baselines are COMPAS or Human scores on their own (1-10 scale, or binarized at $>=5$), or models trained with defendant and Turk worker features and the single score to predict ground truth. All hybrid and single models in this paper were trained using the random forest model class, a model class shown to perform well on many problems (Caruana and Niculescu-Mizil 2006). We use area under the ROC curve (AUC) as our main accuracy measure, in line with several papers measuring recidivism (US Sentencing Commission 2004), but we also report other metrics in Appendix D. All metrics are reported over ten 80%-20% train-test splits to account for variability between test sets. See Appendix A for more details on the hybrid models and error metrics.

Analysis and Results

In this section, we report our findings of COMPAS and Human complementarity in terms of predictive performance and decision making, characterize the space of COMPAS and Human agreement and disagreement, and discuss results from our hybrid models.

COMPAS vs. Humans: Predictive Performance

Across 1,000 defendants, Human scores have slightly higher means than COMPAS (mean HNR 5.1, HWR 5.2, COMPAS 4.6, all on 1-10 scale), and the Human scores are more correlated with each other than with COMPAS (COMPAS and HNR correlation 0.52, COMPAS and HWR 0.53, HNR and HWR 0.93).

Table 2 displays the predictive performance of COMPAS and Human scores, on all defendants and by race (this is similar to Table 1 in Dressel and Farid (2018)). Most scores achieved accuracies around 0.66 and AUCs around 0.70 when evaluated on all races, blacks, or whites. All scores performed slightly worse for other races at approximately 0.65; there are only a small number of these defendants in the data (9%). These findings replicate Dressel and Farid's findings when evaluating accuracies of the three scores.

Table 3 presents predictive performance separated by recidivism status: whether the defendant recidivated or not. Here, Humans were better at predicting defendants who recidivate, while COMPAS was better at predicting defendants who do not recidivate. In other words, on this data, Humans tended to have higher true positive rates (and hence lower false negative rates) and COMPAS tended to have higher true negative rates (and hence lower false positive rates).

Table 2: COMPAS and Human accuracy and AUC when predicting ground truth recidivism.

	Accuracy			AUC		
Race	C	HNR	HWR	C	HNR	HWR
All	0.65	**0.66**	**0.66**	0.70	0.71	0.71
Black	**0.68**	0.66	0.65	0.70	0.69	0.69
White	**0.66**	**0.66**	0.64	0.71	0.69	0.70
Other	0.65	0.60	**0.66**	0.64	0.65	0.65

Table 3: Refinement of Table 2 by recidivism status. Left: defendants who **do** recidivate. Right: defendants who **do not** recidivate. Only accuracies are displayed because AUC cannot be calculated when ground truth only has one value ("yes" for do recidivate, "no" for do not recidivate).

	Accuracy					
	Do recidivate			Do not recidivate		
Race	C	HNR	HWR	C	HNR	HWR
All	0.62	0.68	**0.69**	**0.69**	0.64	0.63
Black	**0.74**	**0.74**	0.70	0.61	0.55	0.59
White	0.60	0.50	0.59	0.69	**0.75**	0.68
Other	0.38	0.59	0.65	**0.80**	0.61	0.66

We see similar effects for the level of agreement between risk scores, race, and ground truth. COMPAS and Humans demonstrate higher levels of agreement for correctly predicting that black defendants will recidivate, but their level of agreement drops significantly for white or other race defendants who recidivate. The opposite is true for defendants who do not recidivate. COMPAS and Humans have higher levels of agreement for correctly predicting that white and other race defendants will not recidivate, but this level of agreement drops for black defendants who do not recidivate.

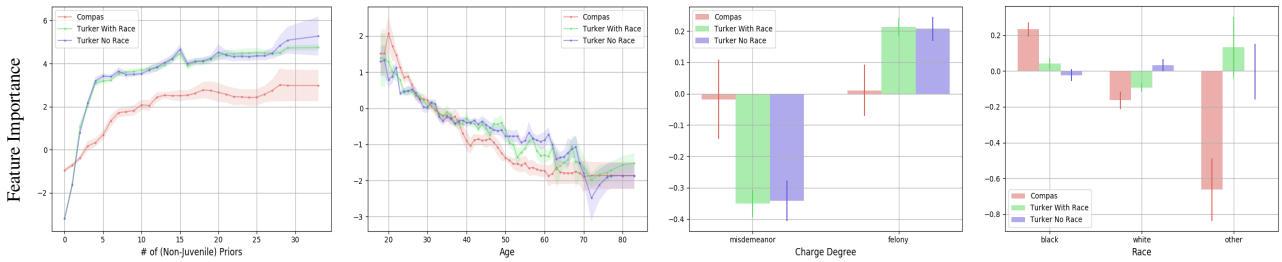

Figure 1: Predicting COMPAS (red), HWR (green), and HNR (purple) scores from features. The larger the y-axis magnitude, the more important the feature. "Number of priors", with y-axis scale -3 to 6, is the most important feature for all three scores, followed by "age".

COMPAS vs. Humans: Decision Making

Which features are most important in COMPAS and Human decision making? It is known that COMPAS scores can be predicted from only a few features, in particular the "number of priors" and age (Chouldechova and G'Sell 2017; Angelino et al. 2017; Tan et al. 2018). To determine if Human decision making places more importance on other features, we trained interpretable models to predict each of the three scores. All three models saw the same set of features – age, race, sex, number of juvenile misdemeanors, number of juvenile felonies, number of (non-juvenile) priors, crime charge degree (misdemeanor or felony), and crime charge. First, we trained iGAM models, a type of additive model based on nonparametric base learners (Caruana et al. 2015). Figure 1 illustrates the importance of four of these features for predicting each score. **Like COMPAS, the two most important features in Human decision making are the "number of priors" and "age".** However, Human scores place more weight on the "number of priors" and "charge degree" features than COMPAS, whereas age's impact is similar for COMPAS and Human scores. Decision trees trained to predict each of the three risk scores confirm that "number of priors" is the most important feature, with every tree's root node splitting on this feature.

Including race when predicting these scores, even when the scores may not have seen race, returns some interesting findings. Recall that HNR scores were generated from Turk workers who were not told the defendants' race. We considered the impact of race on Human recidivism predictions, by comparing the importance of the race feature on HWR (green) and HNR (purple) scores in Figure 1. We find that black defendants were assessed to have slightly higher recidivism risk by Turk workers when told of their race. The decision tree predicting the difference of HWR and HNR scores in Figure A4 also agreed with this finding, returning a first split on race where white defendants were assigned slightly lower risk (-0.16) in the Human with-race condition, and black and other race individuals were assigned slightly higher risk (+0.14). In contrast, both decision trees predicting the difference between COMPAS and HWR scores, as well as COMPAS and HNR scores, split on "number of priors" and age but not race.

Hence, even though revealing race did not significantly affect the predictive performance of Humans for ground truth, as found by Dressel and Farid, including race appeared to have slightly affected Humans' perception of recidivism risk (magnitude around +/- 0.15 on a 1-10 score scale). Note, however, that the set of Turk workers in the no-race and with-race conditions were different; this effect may diminish or exacerbate if the experiment is re-run with the same set of Turk workers.

COMPAS + Humans: Characterizing Agreement & Disagreement

We now determine the features that drive agreement or disagreement between COMPAS and Human scores. To do so, we use two techniques – clustering and decision trees. Specifically, we performed mean-shift clustering (Derpanis 2005), a robust-clustering method that avoids the need to specify an arbitrary number of clusters, to cluster defendants in each of Cases 1-8 from Table 1. We also built a multiclass decision tree to classify individuals into each of the eight cases. Finally, we assessed the distribution of features across the found clusters and tree partitions. Figure A3 presents the decision tree. We elaborate on our findings below. A summary of the feature characteristics is in Table 1.

Easy calls: COMPAS and Humans agree, both correct. When we cluster defendants in this region of correct agreement, two clusters emerge that map to the two cases. The key separation between Cases 1 (COMPAS high, human high, both correct) and 2 (COMPAS low, Human low, both correct) is the number of priors, and to a lesser extent age. The average number of priors for defendants in Case 1 is 7.9, and 0.34 for Case 2. The average age for defendants in Case 1 is 30.3, and 40.56 for Case 2. Consequently, these cases correspond to what one might consider *easy calls*, i.e., defendants for whom the number of priors and age alone provide sufficient information to predict recidivism accurately.

Unknown unknowns: COMPAS and Humans agree, but both incorrect. Now we turn our attention to the region of wrong agreement - defendants whose COMPAS and Human scores agree, yet fail to predict ground truth (Cases 7 & 8). These defendants are very similar to defendants in other cases – they are truly *unknown unknowns*. Effectively, defendants in Cases 7 & 8 are exactly defendants

for whom the number of priors and age alone are not different enough to distinguish them from defendants in other cases, despite these defendants having fundamentally different ground truth labels. Because both COMPAS and Human scores are over reliant on the number of priors and age, both scores fail for defendants for whom these two features alone are not sufficient to predict recidivism.

Characterizing the space of disagreement. Our key finding for defendants for whom COMPAS and Human scores disagree (Cases 3-6) mirrors our findings for the unknown unknowns. These defendants had similar number of priors and age as defendants in other cases. In general, the four cases in the space of disagreement could not be cleanly separated from each other – Cases 3 and 6 were similar; Cases 4 and 5 were similar – and also overlapped with the space of agreement. For example, defendants with low COMPAS scores, high Human scores but did not recidivate (Case 4) tended to have 1.5 to 5.5 priors and are younger than 32.5 years old. However, these defendants significantly overlap with defendants in several other cases (Cases 1, 7, and 5 as seen in Table 1). See Appendix B for details.

COMPAS + Humans: Leveraging Disagreement to Build Hybrid Models

Since defendants for whom COMPAS and Human scores disagree have the highest possibility of benefiting from hybrid models, we build two separate sets of hybrid models: (1) models on all defendants; (2) models on only the space of disagreement (32% of defendants in this data). Table 4 reports the results of hybrid models trained and tested on only these defendants; results for the first set of hybrid models are in Appendix D.

Hybrid methods tended to outperform single scores (or models trained on features and single scores) by a small margin. In Table 4, the best performing model (AUC 0.60) is a hybrid random forest predicting ground truth using features, COMPAS, and Human (no-race condition) scores. This was better than single risk scores (HNR 0.56, HWR 0.54, Compas 0.49), but comparable to a random forest model trained on the original features plus the HNR scores (but not with COMPAS), which obtained an AUC of 0.59. Interestingly, despite the low AUC of COMPAS (0.49), combining it with HNR did not degrade the hybrid model's performance and in fact led to a small AUC improvement of 0.01. However, this improvement is within the margin of error.

Next, we examine these results by race. Table A2 presents these results for blacks, Table A3 for whites, and Table A4 for other races. The trend is again similar, where hybrid models tended to obtain slightly better results than their single-model counterparts, but improvements are typically within the margin of error. Hybrid models for blacks had the best accuracy and error rates; single models for other races (only 31 defendants) had the best accuracy and error rates.

In general, as can be seen in Table A1 to A4, the best hybrid models tended to leverage defendant and Human worker features, plus both risk scores, to either directly or indirectly predict ground truth recidivism. For the space of disagreement, the best hybrid models also tended to prefer HNR over HWR, particularly when evaluating races other than whites. On the other hand, for the space of disagreement, hybrid models based on (weighted) averages of the COMPAS and human scores tend to underperform models that incorporated defendant and Human worker features. Notably, this is not the case for all defendants as the best performing hybrid models for all defendants were the optimally weighted average models (Table A6).

Table 4: Test-set performance of hybrid models built on individuals whose COMPAS and Human risk scores disagree. Best results in cyan and bolded. See Table A1 in the appendix for extended version of this table.

Type	Model	AUC
Hybrid	Best hybrid of C and HNR	**0.60 ± 0.07**
	Best hybrid of C and HWR	0.58 ± 0.08
	Best hybrid of C, HWR, HNR	0.58 ± 0.07
Single	Predict GT from features and HNR	0.59 ± 0.07
	HNR (1-10 scale)	0.56 ± 0.05
	Predict GT from features and HWR	0.54 ± 0.06
	HWR (1-10 scale)	0.54 ± 0.04
	Predict GT from features and C	0.51 ± 0.07
	C (1-10 scale)	0.49 ± 0.06
None	Predict GT from features	0.52 ± 0.07
Random	Randomly pick between C and HNR	0.55 ± 0.08
	Randomly pick between C and HWR	0.54 ± 0.07
	Randomly pick between C, HWR, HNR	0.54 ± 0.06

We have shown that for defendants for whom COMPAS and Human scores disagree, hybrid models can be more beneficial than single risk scores (even when one of the scores is not as high performing as the other, as is the case for COMPAS compared to Humans for this set of individuals), but, in general, the improvements are marginal and, in many cases, within the margin of error.

Discussion

Our key finding is that **on this data set, COMPAS and Human decision making differed, but not in ways that could be leveraged to improve ground truth prediction.** From our analysis, the number of priors is a key feature in both COMPAS and Human decision making. We saw that COMPAS and Humans tended to agree (and were right) on defendants with a very high or very low number of priors. We saw that the defendants that COMPAS and humans agreed on (but were wrong) were truly *unknown unknowns* – there was no discernable pattern in these cases. Unfortunately, they make up 19% of the data, which bounds the maximal possible improvement from a hybrid model on this data.

When we focused on the 32% of defendants where COMPAS and Human decisions disagree, our hybrid models started to exhibit some improvement, though still within the margin of error. The cases in this region were also the most uncertain, with single risk scores achieving between 0.49 and 0.56 AUC. We saw that for this region of uncertainty, single risk scores could be further improved by allowing them to see some amount of ground truth labels, along-

side defendant features. We saw that number of priors, once again, and age were the two most important features to determine whether a defendant would fall in Case 3-6, although separation between these four cases was often not clear.

Several reasons could explain why we were not getting better accuracy from the hybrid models: 1) Ground truth labels are noisy. 2) Turk workers are not experts. 3) Ground truth is inherently unpredictable or the features we have do not present enough information to predict ground truth accurately. 4) Small sample size.

Noisy ground truth labels

One limitation of our hybrid models is possible noise (or bias) in the ground truth labels in the ProPublica COMPAS data. The "primary" definition of recidivism from the US Sentencing Commission (2004) is one of the following during the defendant's initial two years back in the community: (1) re-conviction for a new offence; (2) re-arrest with no conviction; (3) supervision revocation. Although this definition has traditionally been considered reliable, it is only a proxy for ground truth and does not cover defendants arrested but not convicted, or defendants not arrested despite committing crimes. Use of this definition is also susceptible to racial or socioeconomic bias, as people of color or those who live in poorer communities may experience higher levels of policing, resulting in a higher rates of re-arrests (Eckhouse 2017). As we continue to develop machine learning models for recidivism, we need to reevaluate the ground truth labels we are collecting to ensure they are unbiased.

Criminal justice expertise

It is important to note that the Human risk scores in our analyses were obtained from Mechanical Turk workers. The ecological validity of using Turk workers may be low, as they have no criminal justice experience, and the decisions they are asked to make (whether a defendant recidivates or not) may have little relation to the types of decisions they make in their day-to-day lives. Gathering human data from judges, in actual legal settings, will help us further investigate the potential of hybrid models in fairness domains. We need to gather more quantitative and qualitative data on when judges and algorithmic systems agree and disagree, and what additional information the judge may be using to inform her decision. This could help hybrid models better discern when to choose human judgment over algorithmic prediction to achieve better performance overall.

Lacking evidence about the world

We have two other hypotheses why our hybrid models only marginally improve over the accuracy of COMPAS or Human scores alone despite the presence of differences in COMPAS and Human reasoning. First, perhaps recidivism is an unpredictable event with a lot of inherent uncertainty, and as such, the accuracy of any model is limited. This is consistent with prior work that found similar AUCs for commercial recidivism prediction systems (Drake 2014). Second, it could be that the seven features included in this data are not sufficient to properly evaluate recidivism risk. This second explanation is likely, since the Turk worker ratings are only based on those seven features (besides race). In a real world court setting, a judge has access to additional information that could be used to inform their reasoning. This "private information" may be helpful, however it remains to be seen if private information may also be detrimental to human reasoning, as seen in Grove et al. (2000) and Kleinberg et al. (2017) where physicians and judges sometimes responded to private information in ways that caused them to deviate from optimal judgment. In general, tracking as many features as possible would enable more detailed study of the value of private information in complex decision making settings. Moreover, our analyses found that both COMPAS and Human both relied heavily on the number of priors and age. While these features may be considered "objective" (e.g. prior research showed a strong correlation between prior criminal record and recidivism (United States Sentencing Commission 2016)), many defendants may appear similar when viewed through the lens of only two features.

Small sample size

In our hybrid models trained on only the 340 defendants for which COMPAS and Human scores disagreed, the improvements demonstrated were subsumed by large margin of errors. This was also the case for further subgroups of races (169 blacks, 114 whites, 31 other races). Repeating the Mechanical Turk experiment and hybrid models on a larger sample of the original ProPublica COMPAS data will provide more evidence as to whether human judgment can help machines in making recidivism predictions.

Conclusion

In complex settings, like a courtroom or hospital, it is unlikely that algorithmic systems will be making all decisions without input from human experts. Our approach focused efforts on cases where humans and machines disagree as a potential area to enhance decision making. Ultimately, we want to leverage the best of both worlds: humans that glean subtle, interpersonal insights from rich context, and machine algorithms that provide rigor and consistency. However, on this data set, human and machine decision making differed but not in ways that could be leveraged to improve ground truth prediction of recidivism. An important next step will be to further our investigation to include predictions made by judges in real-world settings. We hypothesize that the richness of the real-world may provide better context for enhanced hybrid Human + Machine models.

A key debate in recidivism predictions involves issues of bias and fairness, particularly for false positive and false negative judgments. Although our work uncovered a few aspects where race had an impact, it was not the primary focus of our work. We intend to look more closely at issues of bias and fairness in future work, especially as we gather more real-world data. Although both humans and algorithms can have inherent biases, if these biases are different, a hybrid model has the potential to help overcome them.

References

Andrews, D. A.; Bonta, J.; and Wormith, J. S. 2006. The recent past and near future of risk and/or need assessment. *Crime & Delinquency*.

Angelino, E.; Larus-Stone, N.; Alabi, D.; Seltzer, M.; and Rudin, C. 2017. Learning certifiably optimal rule lists. In *KDD*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica.

Berk, R. A., and Bleich, J. 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Journal of Criminology and Public Policy*.

Blomberg, T.; Bales, W.; Mann, K.; Meldrum, R.; and Nedelec, J. 2010. Validation of the compas risk assessment classification instrument. *Technical Report, Florida State University*.

Caruana, R., and Niculescu-Mizil, A. 2006. An empirical comparison of supervised learning algorithms. In *ICML*.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*.

Chouldechova, A., and G'Sell, M. 2017. Fairer and more accurate, but for whom? In *FATML Workshop*.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*.

Derpanis, K. G. 2005. Mean shift clustering. *Lecture Notes*.

Drake, E. 2014. Predicting criminal recidivism: A systematic review of offender risk assessments in washington state. *Technical Report, Washington State Institute for Public Policy*.

Dressel, J., and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*. Data at `www.cs.dartmouth.edu/farid/downloads/publications/scienceadvances17`.

Eckhouse, L. 2017. Big data may be reinforcing racial bias in the criminal justice system. Washington Post.

Gendreau, P.; Freeze, T.; and Goggin, C. 1996. A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*.

Grgić-Hlača, N.; Redmiles, E. M.; Gummadi, K. P.; and Weller, A. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *WWW*.

Grove, W. M.; Zald, D. H.; Lebow, B. S.; and Nelson, B. E. S. C. 2000. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*.

Horvitz, E., and Paek, T. 2007. Complementary computing: policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction*.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on Amazon Mechanical Turk. In *KDD Workshop on Human Computation*.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics*.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent trade-offs in the fair determination of risk scores. In *ITCS*.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *IUI*.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Horvitz, E. 2017a. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *AAAI*.

Lakkaraju, H.; Kleinberg, J.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017b. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *KDD*.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the compas recidivism algorithm. ProPublica.

Nushi, B.; Kamar, E.; and Horvitz, E. 2018. Towards Accountable AI: Hybrid human-machine analyses for characterizing system failure. In *HCOMP*.

Phillips, P. J.; Yates, A. N.; Hu, Y.; Hahn, C. A.; Noyes, E.; Jackson, K.; Cavazos, J. G.; Jeckeln, G.; Ranjan, R.; Sankaranarayanan, S.; Chen, J.-C.; Castillo, C. D.; Chellappa, R.; White, D.; and O'Toole, A. J. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *PNAS*.

Tan, S.; Caruana, R.; Hooker, G.; and Lou, Y. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *AAAI/ACM AIES*.

United States Sentencing Commission. 2016. Recidivism among federal offenders: A comprehensive overview.

US Sentencing Commission. 2004. Measuring recidivism: the criminal history computation of the federal sentencing guidelines.

Wang, D.; Khosla, A.; Gargeya, R.; Irshad, H.; and Beck, A. H. 2016. Deep learning for identifying metastatic breast cancer. *CoRR* abs/1606.05718.

Zeng, J.; Ustun, B.; and Rudin, C. 2016. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society*.

Appendix A: Details of Hybrid Models and Metrics

Direct and indirect hybrid models. If we had access to an oracle that can be queried to obtain ground truth recidivism for any new observation, we can determine which of COMPAS or Human scores better predicts ground truth. However, test-time access to a ground truth oracle is not realistic. Hence, we relax this assumption of oracle access at test-time to only training-time, and train a binary classification model **only on observations where the two risk scores disagree** to predict which risk score to pick. In other words, this model predicts which score – COMPAS or Human – to use for Cases 3-6 in Table 1 using features available at training time such as defendant features, Turk worker features, COMPAS score, and Human score. We call this an **indirect** hybrid model – indirect because the hybrid model takes as input the prediction of which risk score is better, and outputs the desired ground truth recidivism prediction. Figure A1 shows this model. We also **directly** predict ground truth recidivism as a function of not just features but also the two risk scores. Figure A2 shows this model.

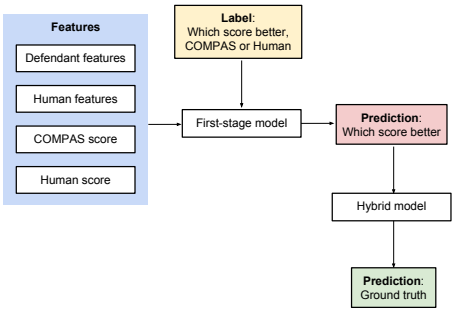

Figure A1: Indirect hybrid model.

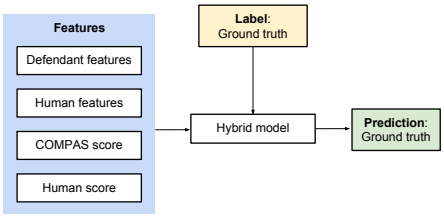

Figure A2: Direct hybrid model.

Accuracy and error metrics. Besides AUC, we also report balanced accuracy (Bal Acc), i.e., the mean classification accuracy across classes. For error rates, we track false positives (FPR), false negatives (FNR), false discovery (FDR), and false omission (FOR). Equations for these error rates are below. Note that Kleinberg (Kleinberg, Mullainathan, and Raghavan 2017) and Choudechouva (Chouldechova 2017) showed the impossibility of satisfying several of these metrics simultaneously.

Given a binary label and a binary prediction, let FP denote the number of false positives, FN denote the number of false negatives, TP denote the number of true positives, and TN denote the number of true negatives.

Balanced accuracy

$$BalAcc = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

False positive rate (FPR)

$$FPR = \frac{FP}{FP+TN}$$

False negative rate (FPR)

$$FPR = \frac{FN}{FN+TP}$$

False discovery rate (FDR)

$$FDR = \frac{FP}{FP+TP}$$

False omission rate (FOR)

$$FOR = \frac{FN}{FN+TN}$$

Appendix B: Detailed Characterization of the Space of Disagreement

COMPAS score high, Human score low (Cases 3 & 6). The difference between Cases 3 and 6 is their ground truth label - defendants in Case 3 recidivated, whereas defendants in Case 6 did not. According to the decision tree's partitions, defendants in Cases 3 and 6 tend to have < 0.5 priors. In fact, the key distinguishing feature between Cases 3 and 6 is the type of crime that the defendant was charged with. In addition, we found that some of the multiclass trees we built to predict classification into the eight cases did not always have terminal nodes with Case 6. Sometimes, Case 6 is combined with Case 3, indicating that the features do not have sufficient signal to adequately distinguish these two cases.

COMPAS score low, Human score high (Cases 4 & 5). The difference between Cases 4 and 5 is also their ground truth label - defendants in Case 4 did not recidivate, whereas defendants in Case 5 did. Case 4 defendants tended to have 1.5 to 5.5 priors and be older than 32.5 years old. Case 5 was not always present as a terminal node in our trees, and are very similar to defendants in Case 4 and also Cases 1 and 7 (in the space of agreement).

Appendix B: Additional Figures

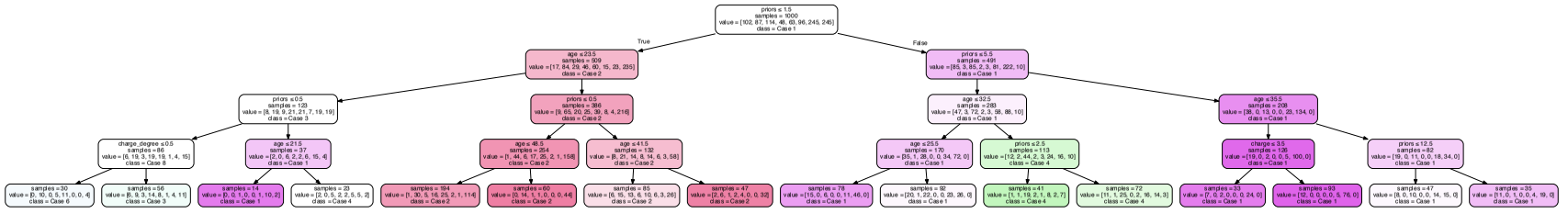

Figure A3: Decision tree to explain the three-way interaction between COMPAS, Human scores, and ground truth. The label for the prediction task corresponds to the 8 different cases from Table 1. The five values in each node are (1) the partition of features defining that node (2) the number of samples in that node (3) the number of samples in each of the 8 cases (4) the predicted case (of the 8 cases). Best viewed after downloading and zooming in a PDF reader.

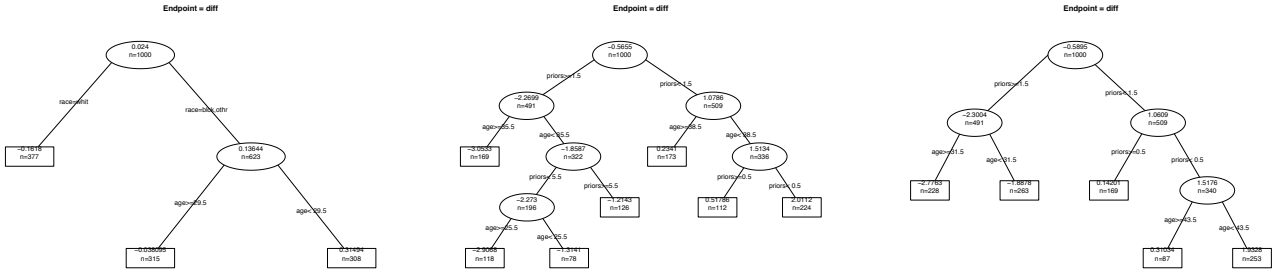

Figure A4: Decision tree predicting the difference between scores. Left: difference in scores given by Turk workers when and when not told of the defendant's race (HWR - HNR). Center: difference in scores given by COMPAS and Turk workers not told of the defendant's race (C - HNR). Right:difference in scores given by COMPAS and Turk workers told of the defendant's race (C - HWR).

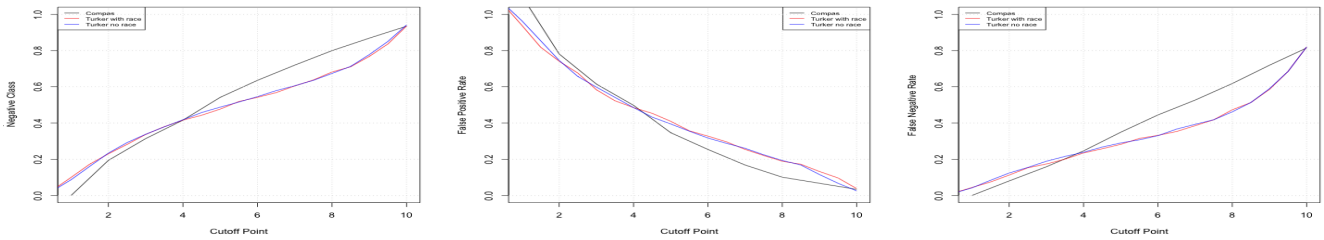

Figure A5: Accuracies (left), false positive rates (center), and false negative rates (right) for COMPAS and Human scores at different cutoff points for binarizing the scores.

Appendix C: Extended result tables for hybrid models for defendants whose COMPAS and Human scores disagree

Table A1: Test-set performance of hybrid models built on individuals whose COMPAS and Human risk scores disagree, compared to just using a single risk score and other baselines. The numbers presented are means and standard deviations calculated over 10 train-test splits. Best results in cyan and bolded. A reduced version of this table can be seen in Table 4. Rows marked with * are the rows labeled as *best* in Table 4.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Hybrid	Direct C HNR*	**0.60 ± 0.07**	**0.56 ± 0.07**	0.44 ± 0.13	0.45 ± 0.10	0.50 ± 0.10	**0.39 ± 0.08**
Hybrid	Composed indirect C HWR*	0.58 ± 0.08	**0.56 ± 0.08**	0.37 ± 0.10	0.50 ± 0.10	**0.47 ± 0.15**	0.40 ± 0.10
Hybrid	Direct C HWR HNR*	0.58 ± 0.07	0.55 ± 0.08	0.47 ± 0.14	0.43 ± 0.09	0.50 ± 0.09	0.40 ± 0.10
Hybrid	Indirect C HWR*	0.58 ± 0.08	**0.56 ± 0.08**	0.37 ± 0.10	0.50 ± 0.10	**0.47 ± 0.15**	0.40 ± 0.10
Hybrid	Composed indirect C HNR	0.56 ± 0.09	0.54 ± 0.06	0.45 ± 0.07	0.47 ± 0.09	0.52 ± 0.07	0.40 ± 0.08
Hybrid	Indirect C HNR	0.56 ± 0.09	0.54 ± 0.06	0.45 ± 0.07	0.47 ± 0.09	0.52 ± 0.07	0.40 ± 0.08
Hybrid	Direct C HWR	0.53 ± 0.06	0.52 ± 0.04	0.37 ± 0.09	0.58 ± 0.09	0.52 ± 0.14	0.44 ± 0.08
Hybrid	Weighted average of C HNR	0.51 ± 0.05	0.50 ± 0.04	0.38 ± 0.25	0.63 ± 0.3	0.56 ± 0.22	0.43 ± 0.07
Hybrid	Weighted average of C HWR HNR	0.50 ± 0.04	0.50 ± 0.05	**0.23 ± 0.07**	0.77 ± 0.13	0.58 ± 0.09	0.45 ± 0.11
Hybrid	Weighted average of C HWR	0.47 ± 0.04	0.49 ± 0.03	0.39 ± 0.26	0.63 ± 0.26	0.56 ± 0.12	0.46 ± 0.11
Single	Predict GT from features and HNR	0.59 ± 0.07	0.55 ± 0.06	0.44 ± 0.09	0.46 ± 0.10	0.51 ± 0.09	**0.39 ± 0.07**
Single	HNR (1-10 scale)	0.56 ± 0.05	0.52 ± 0.02	0.55 ± 0.08	0.40 ± 0.08	0.54 ± 0.04	0.41 ± 0.07
Single	Predict GT from features and HWR	0.54 ± 0.06	0.54 ± 0.05	0.35 ± 0.10	0.57 ± 0.08	0.49 ± 0.14	0.42 ± 0.09
Single	HWR (1-10 scale)	0.54 ± 0.04	0.52 ± 0.03	0.54 ± 0.05	0.41 ± 0.04	0.53 ± 0.09	0.43 ± 0.10
Single	Predict GT from features and C	0.51 ± 0.07	0.52 ± 0.05	0.41 ± 0.07	0.55 ± 0.08	0.52 ± 0.11	0.44 ± 0.10
Single	C (1-10 scale)	0.49 ± 0.06	0.48 ± 0.01	0.40 ± 0.07	0.65 ± 0.08	0.59 ± 0.06	0.46 ± 0.04
Single	C (binarized >=5)	-	0.48 ± 0.01	0.40 ± 0.07	0.65 ± 0.08	0.59 ± 0.06	0.46 ± 0.04
Single	HNR (binarized >=5)	-	0.52 ± 0.01	0.60 ± 0.07	**0.35 ± 0.08**	0.54 ± 0.04	0.41 ± 0.06
Single	HWR (binarized >=5)	-	0.51 ± 0.03	0.63 ± 0.05	0.36 ± 0.05	0.54 ± 0.07	0.44 ± 0.12
None	Predict GT from features	0.52 ± 0.07	0.51 ± 0.06	0.37 ± 0.07	0.61 ± 0.09	0.54 ± 0.13	0.45 ± 0.09
Random	Randomly pick between C HNR	0.55 ± 0.08	0.52 ± 0.05	0.46 ± 0.06	0.50 ± 0.08	0.54 ± 0.07	0.42 ± 0.07
Random	Randomly pick between C HWR	0.54 ± 0.07	0.52 ± 0.06	0.46 ± 0.06	0.49 ± 0.11	0.53 ± 0.14	0.43 ± 0.08
Random	Randomly pick between C HWR HNR	0.54 ± 0.06	0.52 ± 0.06	0.50 ± 0.07	0.46 ± 0.08	0.53 ± 0.10	0.43 ± 0.11

Table A2: Subgroup (African-Americans) performance of models presented in Table A1.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Hybrid	Direct C HNR	**0.65 ± 0.06**	0.58 ± 0.07	0.48 ± 0.15	0.35 ± 0.13	0.46 ± 0.11	0.37 ± 0.10
Hybrid	Direct C HWR HNR	0.63 ± 0.07	**0.59 ± 0.07**	0.50 ± 0.15	**0.32 ± 0.11**	0.45 ± 0.09	**0.36 ± 0.13**
Hybrid	Composed indirect C HWR	0.57 ± 0.12	0.58 ± 0.10	0.41 ± 0.15	0.43 ± 0.14	0.43 ± 0.17	0.41 ± 0.12
Hybrid	Indirect C HWR	0.57 ± 0.12	0.58 ± 0.10	0.41 ± 0.15	0.43 ± 0.14	0.43 ± 0.17	0.41 ± 0.12
Hybrid	Weighted average of C HNR	0.55 ± 0.06	0.51 ± 0.07	0.33 ± 0.19	0.64 ± 0.31	0.65 ± 0.21	0.42 ± 0.12
Hybrid	Weighted average of C HWR HNR	0.55 ± 0.08	0.55 ± 0.08	**0.12 ± 0.09**	0.78 ± 0.15	**0.35 ± 0.21**	0.45 ± 0.15
Hybrid	Composed indirect C HNR	0.53 ± 0.09	0.52 ± 0.07	0.54 ± 0.10	0.41 ± 0.07	0.52 ± 0.07	0.44 ± 0.13
Hybrid	Indirect C HNR	0.53 ± 0.09	0.52 ± 0.07	0.54 ± 0.10	0.41 ± 0.07	0.52 ± 0.07	0.44 ± 0.13
Hybrid	Direct C HWR	0.51 ± 0.07	0.51 ± 0.07	0.40 ± 0.14	0.57 ± 0.12	0.50 ± 0.17	0.48 ± 0.10
Hybrid	Weighted average of C HWR	0.48 ± 0.09	0.49 ± 0.07	0.37 ± 0.27	0.65 ± 0.26	0.50 ± 0.16	0.51 ± 0.15
Single	Predict GT from features and HNR	0.64 ± 0.06	0.56 ± 0.06	0.48 ± 0.13	0.39 ± 0.13	0.48 ± 0.10	0.39 ± 0.09
Single	HNR (1-10 scale)	0.55 ± 0.07	0.56 ± 0.05	0.46 ± 0.09	0.42 ± 0.15	0.49 ± 0.08	0.39 ± 0.09
Single	HWR (1-10 scale)	0.53 ± 0.08	0.53 ± 0.06	0.47 ± 0.08	0.47 ± 0.11	0.49 ± 0.13	0.46 ± 0.13
Single	Predict GT from features and HWR	0.52 ± 0.08	0.53 ± 0.08	0.36 ± 0.15	0.57 ± 0.13	0.46 ± 0.17	0.46 ± 0.12
Single	Predict GT from features and C	0.49 ± 0.11	0.50 ± 0.06	0.48 ± 0.12	0.52 ± 0.13	0.52 ± 0.13	0.49 ± 0.12
Single	C (1-10 scale)	0.46 ± 0.06	0.44 ± 0.05	0.51 ± 0.10	0.60 ± 0.16	0.61 ± 0.09	0.51 ± 0.08
Single	C (binarized >=5)	-	0.44 ± 0.05	0.51 ± 0.10	0.60 ± 0.16	0.61 ± 0.09	0.51 ± 0.08
Single	HNR (binarized >=5)	-	0.56 ± 0.05	0.49 ± 0.10	0.40 ± 0.16	0.49 ± 0.08	0.39 ± 0.09
Single	HWR (binarized >=5)	-	0.51 ± 0.06	0.57 ± 0.11	0.42 ± 0.12	0.51 ± 0.11	0.48 ± 0.16
None	Predict GT from features	0.49 ± 0.10	0.48 ± 0.10	0.42 ± 0.13	0.62 ± 0.12	0.54 ± 0.19	0.50 ± 0.11
Random	Randomly pick between C HWR	0.59 ± 0.06	0.57 ± 0.07	0.47 ± 0.08	0.40 ± 0.13	0.46 ± 0.14	0.41 ± 0.11
Random	Randomly pick between C HWR HNR	0.54 ± 0.07	0.53 ± 0.06	0.48 ± 0.09	0.47 ± 0.11	0.49 ± 0.11	0.46 ± 0.13
Random	Randomly pick between C HNR	0.53 ± 0.13	0.50 ± 0.10	0.53 ± 0.11	0.47 ± 0.11	0.54 ± 0.07	0.46 ± 0.15

Table A3: Subgroup (whites) performance of models presented in Table A1.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Hybrid	Composed indirect C HWR	0.59 ± 0.09	0.55 ± 0.08	**0.30 ± 0.14**	0.61 ± 0.11	0.49 ± 0.21	0.39 ± 0.10
Hybrid	Indirect C HWR	0.59 ± 0.09	0.55 ± 0.08	**0.30 ± 0.14**	0.61 ± 0.11	0.49 ± 0.21	0.39 ± 0.10
Hybrid	Composed indirect C HNR	0.56 ± 0.2	0.52 ± 0.17	0.41 ± 0.16	0.56 ± 0.2	0.55 ± 0.19	0.42 ± 0.18
Hybrid	Direct C HWR	0.56 ± 0.08	**0.58 ± 0.09**	0.31 ± 0.18	0.53 ± 0.17	**0.45 ± 0.24**	**0.36 ± 0.12**
Hybrid	Indirect C HNR	0.56 ± 0.2	0.52 ± 0.17	0.41 ± 0.16	0.56 ± 0.2	0.55 ± 0.19	0.42 ± 0.18
Hybrid	Direct C HNR	0.53 ± 0.17	0.52 ± 0.13	0.39 ± 0.11	0.57 ± 0.21	0.56 ± 0.17	0.42 ± 0.15
Hybrid	Direct C HWR HNR	0.52 ± 0.19	0.50 ± 0.13	0.43 ± 0.14	0.56 ± 0.2	0.57 ± 0.17	0.44 ± 0.14
Hybrid	Weighted average of C HWR	0.48 ± 0.11	0.48 ± 0.06	0.42 ± 0.28	0.61 ± 0.25	0.59 ± 0.19	0.45 ± 0.14
Hybrid	Weighted average of C HWR HNR	0.46 ± 0.08	0.44 ± 0.05	0.36 ± 0.10	0.76 ± 0.19	0.72 ± 0.13	0.46 ± 0.11
Hybrid	Weighted average of C HNR	0.44 ± 0.08	0.46 ± 0.06	0.47 ± 0.33	0.60 ± 0.3	0.55 ± 0.21	0.51 ± 0.21
Single	Predict GT from features and HWR	0.59 ± 0.09	**0.58 ± 0.08**	0.32 ± 0.15	0.52 ± 0.12	0.46 ± 0.21	**0.36 ± 0.1**
Single	Predict GT from features and C	0.56 ± 0.14	0.55 ± 0.12	0.34 ± 0.10	0.56 ± 0.23	0.53 ± 0.16	0.39 ± 0.14
Single	HWR (1-10 scale)	0.56 ± 0.12	0.53 ± 0.11	0.60 ± 0.18	0.34 ± 0.18	0.55 ± 0.12	0.39 ± 0.22
Single	Predict GT from features and HNR	0.53 ± 0.19	0.53 ± 0.15	0.41 ± 0.12	0.54 ± 0.22	0.55 ± 0.17	0.41 ± 0.17
Single	HNR (1-10 scale)	0.53 ± 0.15	0.46 ± 0.10	0.67 ± 0.15	0.42 ± 0.16	0.59 ± 0.09	0.51 ± 0.25
Single	C (1-10 scale)	0.52 ± 0.11	0.48 ± 0.09	0.35 ± 0.15	0.69 ± 0.17	0.60 ± 0.21	0.44 ± 0.10
Single	C (binarized >=5)	-	0.48 ± 0.09	0.35 ± 0.15	0.69 ± 0.17	0.60 ± 0.21	0.44 ± 0.10
Single	HNR (binarized >=5)	-	0.47 ± 0.07	0.72 ± 0.13	0.34 ± 0.14	0.58 ± 0.08	0.50 ± 0.24
Single	HWR (binarized >=5)	-	0.52 ± 0.09	0.65 ± 0.15	**0.31 ± 0.17**	0.56 ± 0.10	0.40 ± 0.21
None	Predict GT from features	0.55 ± 0.14	0.55 ± 0.12	0.37 ± 0.11	0.53 ± 0.2	0.51 ± 0.16	0.39 ± 0.14
Random	Randomly pick between C HNR	**0.61 ± 0.14**	0.57 ± 0.13	0.37 ± 0.11	0.49 ± 0.18	0.49 ± 0.14	0.38 ± 0.17
Random	Randomly pick between C HWR HNR	0.54 ± 0.12	0.50 ± 0.14	0.54 ± 0.13	0.45 ± 0.18	0.57 ± 0.14	0.42 ± 0.17
Random	Randomly pick between C HWR	0.49 ± 0.08	0.47 ± 0.08	0.47 ± 0.12	0.58 ± 0.13	0.61 ± 0.17	0.45 ± 0.09

Table A4: Subgroup (other races) performance of models presented in Table A1.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Hybrid	Weighted average of C HNR	0.64 ± 0.33	0.58 ± 0.23	0.47 ± 0.33	0.38 ± 0.44	0.63 ± 0.34	0.22 ± 0.25
Hybrid	Composed indirect C HNR	0.62 ± 0.32	0.59 ± 0.24	0.29 ± 0.26	0.53 ± 0.39	0.52 ± 0.38	0.29 ± 0.21
Hybrid	Indirect C HNR	0.62 ± 0.32	0.59 ± 0.24	0.29 ± 0.26	0.53 ± 0.39	0.52 ± 0.38	0.29 ± 0.21
Hybrid	Weighted average of C HWR HNR	0.50 ± 0.25	0.57 ± 0.21	0.58 ± 0.37	0.29 ± 0.3	0.52 ± 0.27	0.33 ± 0.37
Hybrid	Weighted average of C HWR	0.49 ± 0.24	0.52 ± 0.14	0.64 ± 0.26	0.32 ± 0.39	0.59 ± 0.23	0.29 ± 0.37
Hybrid	Direct C HWR HNR	0.46 ± 0.22	0.44 ± 0.22	0.47 ± 0.31	0.66 ± 0.35	0.74 ± 0.33	0.48 ± 0.29
Hybrid	Direct C HNR	0.43 ± 0.22	0.48 ± 0.18	0.32 ± 0.28	0.72 ± 0.25	0.61 ± 0.42	0.39 ± 0.13
Hybrid	Direct C HWR	0.43 ± 0.17	0.39 ± 0.16	0.37 ± 0.23	0.85 ± 0.16	0.72 ± 0.37	0.52 ± 0.15
Hybrid	Composed indirect C HWR	0.39 ± 0.24	0.44 ± 0.18	0.41 ± 0.31	0.70 ± 0.31	0.68 ± 0.37	0.51 ± 0.25
Hybrid	Indirect C HWR	0.39 ± 0.24	0.44 ± 0.18	0.41 ± 0.31	0.70 ± 0.31	0.68 ± 0.37	0.51 ± 0.25
Single	HNR (1-10 scale)	**0.65 ± 0.22**	0.59 ± 0.16	0.60 ± 0.17	0.21 ± 0.25	0.58 ± 0.2	0.25 ± 0.27
Single	Predict GT from features and HWR	0.50 ± 0.18	0.47 ± 0.16	0.3 ± 0.22	0.75 ± 0.19	0.57 ± 0.36	0.47 ± 0.18
Single	Predict GT from features and HNR	0.47 ± 0.26	0.47 ± 0.19	0.31 ± 0.25	0.75 ± 0.27	0.67 ± 0.44	0.38 ± 0.11
Single	HWR (1-10 scale)	0.44 ± 0.26	0.43 ± 0.22	0.71 ± 0.23	0.44 ± 0.28	0.59 ± 0.24	0.52 ± 0.34
Single	Predict GT from features and C	0.36 ± 0.31	0.49 ± 0.2	0.3 ± 0.18	0.71 ± 0.37	0.67 ± 0.37	0.35 ± 0.2
Single	C (1-10 scale)	0.33 ± 0.2	0.47 ± 0.14	**0.21 ± 0.16**	0.85 ± 0.17	0.67 ± 0.41	0.50 ± 0.21
Single	C (binarized >=5)	-	0.39 ± 0.17	0.35 ± 0.22	0.88 ± 0.25	0.88 ± 0.25	0.35 ± 0.14
Single	HNR (binarized >=5)	-	**0.61 ± 0.17**	0.65 ± 0.22	**0.12 ± 0.25**	0.65 ± 0.14	**0.12 ± 0.25**
Single	HWR (binarized >=5)	-	0.53 ± 0.14	0.79 ± 0.16	0.15 ± 0.17	**0.50 ± 0.21**	0.33 ± 0.41
None	Predict GT from features	0.48 ± 0.39	0.54 ± 0.26	0.3 ± 0.2	0.62 ± 0.42	0.61 ± 0.42	0.32 ± 0.23
Random	Randomly pick between C HWR HNR	0.36 ± 0.28	0.44 ± 0.2	0.53 ± 0.22	0.59 ± 0.29	0.64 ± 0.25	0.46 ± 0.25
Random	Randomly pick between C HWR	0.32 ± 0.2	0.34 ± 0.14	0.35 ± 0.21	0.96 ± 0.09	0.86 ± 0.38	0.55 ± 0.16
Random	Randomly pick between C HNR	0.3 ± 0.33	0.33 ± 0.24	0.51 ± 0.32	0.83 ± 0.22	0.75 ± 0.35	0.53 ± 0.22

Appendix D: Result tables for hybrid models for all defendants

Table A5: Test-set performance of hybrid models built on all individuals, compared to just using a single risk score and other baselines. The benevolent oracle is the risk score best at predicting ground truth, to provide an upper bound on the accuracy reachable on this data set of any hybrid COMPAS-Human model built on the two risk scores. The adversarial oracle is the risk score **worse** at predicting ground truth, to provide a lower bound. See Table A6 for extended version of this table.

Type	Model	AUC
Oracle	Benevolent oracle	0.85 ± 0.03
	Adversarial oracle	0.57 ± 0.03
Hybrid	Best hybrid of C and HNR	**0.74 ± 0.03**
	Best hybrid of C and HWR	**0.74 ± 0.04**
	Best hybrid of C, HWR, HNR	0.73 ± 0.03
Single	HNR (1-10 scale)	0.72 ± 0.03
	HWR (1-10 scale)	0.72 ± 0.03
	C (1-10 scale)	0.71 ± 0.03
	Predict GT from features and C	0.71 ± 0.03
	Predict GT from features and HWR	0.71 ± 0.03
	Predict GT from features and HNR	0.70± 0.03
None	Predict GT from features	0.69 ± 0.02
Random	Randomly pick between C and HWR	0.73 ± 0.04
	Randomly pick between C and HNR	0.72 ± 0.04
	Randomly pick between C, HWR, HNR	0.71 ± 0.03

Table A6: Test-set performance of hybrid models built on all individuals, compared to just using a single risk score and other baselines. The numbers presented are means and standard deviations calculated over 10 train-test splits. Best results in cyan and bolded. A reduced version of this table can be seen in Table A5. Rows marked with * are the rows labeled as *best* in Table A5.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Oracle	Benevolent oracle	0.85 ± 0.03	0.81 ± 0.02	0.19 ± 0.04	0.19 ± 0.03	0.20 ± 0.04	0.18 ± 0.03
Oracle	Adversarial oracle	0.57 ± 0.03	0.51 ± 0.02	0.50± 0.03	0.49 ± 0.05	0.53 ± 0.03	0.46 ± 0.04
Hybrid	Weighted average of C HNR*	**0.74 ± 0.03**	0.65 ± 0.06	0.41 ± 0.21	0.29 ± 0.1	0.38 ± 0.06	**0.30 ± 0.04**
Hybrid	Weighted average of C HWR*	**0.74 ± 0.04**	0.65 ± 0.06	0.40 ± 0.2	0.29 ± 0.11	0.36 ± 0.06	0.31 ± 0.05
Hybrid	Direct C HWR HNR*	0.73 ± 0.03	0.66 ± 0.03	0.32 ± 0.05	0.36 ± 0.05	0.35 ± 0.05	0.34 ± 0.05
Hybrid	Direct C HNR	0.72 ± 0.04	0.65 ± 0.03	0.34 ± 0.05	0.36 ± 0.06	0.38 ± 0.04	0.32 ± 0.04
Hybrid	Direct C HWR	0.72 ± 0.03	0.65 ± 0.03	0.32 ± 0.06	0.38 ± 0.06	0.35 ± 0.06	0.35 ± 0.05
Single	HNR (1-10 scale)	0.72 ± 0.03	0.66 ± 0.03	0.35 ± 0.04	0.32 ± 0.04	0.37 ± 0.03	**0.30 ± 0.03**
Single	HWR (1-10 scale)	0.72 ± 0.03	0.66 ± 0.02	0.36 ± 0.04	0.31 ± 0.03	0.36 ± 0.04	0.32 ± 0.04
Single	C (1-10 scale)	0.71 ± 0.03	0.65 ± 0.03	0.32 ± 0.03	0.38 ± 0.06	0.37 ± 0.03	0.33 ± 0.04
Single	Predict GT from features and C	0.71 ± 0.03	0.64 ± 0.04	0.35 ± 0.04	0.36 ± 0.06	0.38 ± 0.04	0.33 ± 0.05
Single	Predict GT from features and HWR	0.71 ± 0.03	**0.67 ± 0.03**	**0.31 ± 0.05**	0.36 ± 0.06	**0.34 ± 0.05**	0.33 ± 0.06
Single	Predict GT from features and HNR	0.70± 0.03	0.64 ± 0.02	0.35 ± 0.04	0.37 ± 0.05	0.39 ± 0.03	0.33 ± 0.04
Single	C (binarized >=5)	-	0.65 ± 0.03	0.32 ± 0.03	0.38 ± 0.06	0.37 ± 0.03	0.33 ± 0.04
Single	HNR (binarized >=5)	-	0.66 ± 0.03	0.38 ± 0.04	0.30 ± 0.04	0.38 ± 0.03	**0.30 ± 0.04**
Single	HWR (binarized >=5)	-	0.66 ± 0.03	0.40 ± 0.04	0.28 ± 0.04	0.37 ± 0.04	0.31 ± 0.05
None	Predict GT from features	0.69 ± 0.02	0.63 ± 0.03	0.37 ± 0.05	0.37 ± 0.06	0.40 ± 0.04	0.34 ± 0.04
Random	Randomly pick between C HWR	0.73 ± 0.04	**0.67 ± 0.03**	0.34 ± 0.03	0.32 ± 0.03	0.35 ± 0.05	0.32 ± 0.04
Random	Randomly pick between C HNR	0.72 ± 0.04	0.66 ± 0.04	0.33 ± 0.03	0.34 ± 0.05	0.36 ± 0.04	0.31 ± 0.04
Random	Randomly pick between C HWR HNR	0.71 ± 0.03	**0.67 ± 0.03**	0.35 ± 0.03	0.31 ± 0.05	0.37 ± 0.03	**0.30 ± 0.04**

Table A7: Subgroup (African-Americans) performance of models presented in Table A6.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Oracle	Benevolent oracle	0.85 ± 0.03	0.81 ± 0.03	0.23 ± 0.03	0.15 ± 0.04	0.18 ± 0.03	0.19 ± 0.05
Oracle	Adversarial oracle	0.54 ± 0.07	0.49 ± 0.05	0.60± 0.08	0.42 ± 0.04	0.46 ± 0.05	0.56 ± 0.08
Hybrid	Direct C HNR	**0.73 ± 0.06**	0.65 ± 0.05	0.47 ± 0.09	0.23 ± 0.08	0.34 ± 0.05	0.34 ± 0.1
Hybrid	Weighted average of C HNR	**0.73 ± 0.05**	0.65 ± 0.07	0.48 ± 0.18	0.22 ± 0.09	0.33 ± 0.05	0.30 ± 0.14
Hybrid	Direct C HWR HNR	0.72 ± 0.03	0.65 ± 0.03	0.44 ± 0.05	0.27 ± 0.06	0.30 ± 0.04	0.41 ± 0.07
Hybrid	Weighted average of C HWR	0.72 ± 0.04	0.64 ± 0.06	0.47 ± 0.19	0.25 ± 0.1	0.30 ± 0.05	0.41 ± 0.07
Hybrid	Weighted average of C HWR HNR	0.71 ± 0.05	0.62 ± 0.07	0.59 ± 0.24	**0.17 ± 0.13**	0.36 ± 0.07	**0.23 ± 0.18**
Hybrid	Direct C HWR	0.70± 0.04	0.62 ± 0.02	0.47 ± 0.07	0.29 ± 0.06	0.32 ± 0.05	0.43 ± 0.05
Single	Predict GT from features and HNR	0.71 ± 0.05	0.63 ± 0.04	0.49 ± 0.08	0.24 ± 0.08	0.35 ± 0.05	0.36 ± 0.1
Single	HNR (1-10 scale)	0.71 ± 0.04	**0.68 ± 0.04**	**0.39 ± 0.06**	0.26 ± 0.05	0.30 ± 0.04	0.34 ± 0.07
Single	Predict GT from features and C	0.70± 0.04	0.62 ± 0.04	0.49 ± 0.08	0.27 ± 0.06	0.32 ± 0.05	0.43 ± 0.08
Single	HWR (1-10 scale)	0.70± 0.03	0.65 ± 0.03	0.43 ± 0.05	0.27 ± 0.04	**0.29 ± 0.04**	0.41 ± 0.05
Single	C (1-10 scale)	0.69 ± 0.05	0.63 ± 0.04	0.43 ± 0.06	0.31 ± 0.07	0.34 ± 0.05	0.39 ± 0.07
Single	Predict GT from features and HWR	0.69 ± 0.04	0.64 ± 0.04	0.45 ± 0.07	0.26 ± 0.06	0.30 ± 0.05	0.40 ± 0.07
Single	C (binarized >=5)	-	0.63 ± 0.04	0.43 ± 0.06	0.31 ± 0.07	0.34 ± 0.05	0.39 ± 0.07
Single	HNR (binarized >=5)	-	0.67 ± 0.04	0.41 ± 0.07	0.25 ± 0.05	0.32 ± 0.05	0.34 ± 0.07
Single	HWR (binarized >=5)	-	0.64 ± 0.03	0.48 ± 0.05	0.24 ± 0.05	0.31 ± 0.04	0.39 ± 0.07
None	Predict GT from features	0.69 ± 0.04	0.62 ± 0.04	0.52 ± 0.06	0.23 ± 0.08	0.36 ± 0.05	0.36 ± 0.11
Random	Randomly pick between C HWR	0.72 ± 0.04	0.66 ± 0.03	0.44 ± 0.04	0.24 ± 0.04	**0.29 ± 0.04**	0.38 ± 0.05
Random	Randomly pick between C HNR	0.70± 0.04	0.65 ± 0.04	0.42 ± 0.05	0.27 ± 0.05	0.32 ± 0.04	0.36 ± 0.07
Random	Randomly pick between C HWR HNR	0.70± 0.05	0.66 ± 0.04	0.42 ± 0.07	0.25 ± 0.05	0.32 ± 0.05	0.35 ± 0.08

Table A8: Subgroup (whites) performance of models presented in Table A6.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Oracle	Benevolent oracle	0.84 ± 0.04	0.8 ± 0.03	0.13 ± 0.06	0.28 ± 0.04	0.23 ± 0.07	0.16 ± 0.05
Oracle	Adversarial oracle	0.55 ± 0.05	0.48 ± 0.05	0.42 ± 0.06	0.61 ± 0.09	0.63 ± 0.08	0.4 ± 0.07
Hybrid	Weighted average of C HWR	**0.75 ± 0.05**	0.65 ± 0.06	0.34 ± 0.22	0.37 ± 0.14	0.46 ± 0.1	**0.23 ± 0.1**
Hybrid	Weighted average of C HWR HNR	0.74 ± 0.05	0.57 ± 0.08	0.64 ± 0.36	**0.21 ± 0.22**	0.55 ± 0.09	**0.23 ± 0.31**
Hybrid	Weighted average of C HNR	0.72 ± 0.03	0.62 ± 0.07	0.35 ± 0.23	0.4 ± 0.16	0.46 ± 0.1	0.26 ± 0.11
Hybrid	Direct C HWR	0.70± 0.04	0.62 ± 0.05	0.21 ± 0.07	0.54 ± 0.1	0.43 ± 0.12	0.29 ± 0.06
Hybrid	Direct C HWR HNR	0.70± 0.04	0.63 ± 0.04	0.21 ± 0.05	0.53 ± 0.08	0.44 ± 0.1	0.29 ± 0.06
Hybrid	Direct C HNR	0.67 ± 0.05	0.62 ± 0.04	0.22 ± 0.04	0.55 ± 0.07	0.43 ± 0.1	0.31 ± 0.05
Single	HWR (1-10 scale)	0.74 ± 0.05	**0.67 ± 0.05**	0.29 ± 0.07	0.38 ± 0.06	0.44 ± 0.08	0.24 ± 0.07
Single	HNR (1-10 scale)	0.70± 0.04	0.62 ± 0.05	0.33 ± 0.04	0.43 ± 0.06	0.47 ± 0.05	0.29 ± 0.07
Single	C (1-10 scale)	0.69 ± 0.06	0.63 ± 0.05	0.24 ± 0.05	0.50± 0.12	0.44 ± 0.1	0.29 ± 0.05
Single	Predict GT from features and HWR	0.69 ± 0.05	0.63 ± 0.06	0.2 ± 0.06	0.54 ± 0.09	0.43 ± 0.12	0.29 ± 0.06
Single	Predict GT from features and C	0.67 ± 0.05	0.63 ± 0.03	**0.19 ± 0.05**	0.55 ± 0.05	**0.4 ± 0.09**	0.3 ± 0.05
Single	Predict GT from features and HNR	0.66 ± 0.06	0.61 ± 0.05	0.23 ± 0.07	0.56 ± 0.08	0.45 ± 0.11	0.32 ± 0.05
Single	C (binarized >=5)	-	0.63 ± 0.05	0.24 ± 0.05	0.50± 0.12	0.44 ± 0.1	0.29 ± 0.05
Single	HNR (binarized >=5)	-	0.63 ± 0.05	0.36 ± 0.04	0.38 ± 0.06	0.47 ± 0.05	0.27 ± 0.07
Single	HWR (binarized >=5)	-	0.66 ± 0.04	0.31 ± 0.07	0.36 ± 0.05	0.45 ± 0.07	0.24 ± 0.07
None	Predict GT from features	0.65 ± 0.05	0.60± 0.05	0.22 ± 0.08	0.57 ± 0.08	0.44 ± 0.13	0.32 ± 0.05
Random	Randomly pick between C HWR	0.73 ± 0.04	0.64 ± 0.03	0.26 ± 0.07	0.46 ± 0.05	0.45 ± 0.08	0.27 ± 0.06
Random	Randomly pick between C HNR	0.72 ± 0.06	0.65 ± 0.06	0.25 ± 0.03	0.45 ± 0.09	0.42 ± 0.07	0.28 ± 0.06
Random	Randomly pick between C HWR HNR	0.70± 0.03	0.65 ± 0.05	0.3 ± 0.05	0.4 ± 0.08	0.44 ± 0.07	0.27 ± 0.06

Table A9: Subgroup (other races) performance of models presented in Table A6.

Type	Model	AUC	Bal Acc	FPR	FNR	FDR	FOR
Oracle	Benevolent oracle	0.84 ± 0.12	0.82 ± 0.09	0.09 ± 0.1	0.27 ± 0.17	0.16 ± 0.17	0.14 ± 0.12
Oracle	Adversarial oracle	0.48 ± 0.1	0.46 ± 0.1	0.45 ± 0.14	0.62 ± 0.17	0.70 ± 0.1	0.38 ± 0.17
Hybrid	Weighted average of C HWR HNR	$\mathbf{0.76 \pm 0.15}$	$\mathbf{0.74 \pm 0.18}$	$\mathbf{0.18 \pm 0.12}$	0.33 ± 0.37	$\mathbf{0.41 \pm 0.32}$	$\mathbf{0.14 \pm 0.13}$
Hybrid	Weighted average of C HNR	0.75 ± 0.15	0.68 ± 0.17	0.30 ± 0.27	0.34 ± 0.29	0.44 ± 0.27	0.25 ± 0.28
Hybrid	Direct C HNR	0.69 ± 0.12	0.56 ± 0.17	0.30 ± 0.1	0.58 ± 0.28	0.62 ± 0.24	0.30 ± 0.16
Hybrid	Direct C HWR	0.69 ± 0.15	0.59 ± 0.12	0.21 ± 0.14	0.61 ± 0.24	0.55 ± 0.3	0.29 ± 0.14
Hybrid	Direct C HWR HNR	0.68 ± 0.08	0.58 ± 0.08	0.27 ± 0.09	0.58 ± 0.2	0.58 ± 0.21	0.28 ± 0.15
Hybrid	Weighted average of C HWR	0.66 ± 0.07	0.62 ± 0.07	0.31 ± 0.12	0.44 ± 0.13	0.52 ± 0.17	0.25 ± 0.11
Single	HNR (1-10 scale)	0.73 ± 0.16	0.69 ± 0.14	0.32 ± 0.17	0.31 ± 0.2	0.44 ± 0.19	0.21 ± 0.18
Single	Predict GT from features and C	0.69 ± 0.12	0.56 ± 0.17	0.29 ± 0.08	0.59 ± 0.27	0.60 ± 0.21	0.30 ± 0.15
Single	Predict GT from features and HNR	0.69 ± 0.14	0.60 ± 0.21	0.26 ± 0.13	0.54 ± 0.37	0.55 ± 0.3	0.29 ± 0.19
Single	Predict GT from features and HWR	0.67 ± 0.14	0.65 ± 0.1	0.22 ± 0.12	0.48 ± 0.17	0.45 ± 0.19	0.25 ± 0.13
Single	HWR (1-10 scale)	0.66 ± 0.07	0.61 ± 0.05	0.37 ± 0.1	0.41 ± 0.12	0.54 ± 0.14	0.26 ± 0.1
Single	C (1-10 scale)	0.64 ± 0.11	0.61 ± 0.09	0.20 ± 0.1	0.57 ± 0.14	0.46 ± 0.13	0.28 ± 0.14
Single	C (binarized >=5)	-	0.61 ± 0.09	0.20 ± 0.1	0.57 ± 0.14	0.46 ± 0.13	0.28 ± 0.14
Single	HNR (binarized >=5)	-	0.71 ± 0.14	0.34 ± 0.17	$\mathbf{0.24 \pm 0.2}$	0.43 ± 0.19	0.18 ± 0.17
Single	HWR (binarized >=5)	-	0.64 ± 0.08	0.43 ± 0.11	0.30 ± 0.14	0.54 ± 0.11	0.23 ± 0.12
None	Predict GT from features	0.68 ± 0.16	0.57 ± 0.21	0.32 ± 0.14	0.53 ± 0.35	0.59 ± 0.28	0.30 ± 0.19
Random	Randomly pick between C HWR HNR	0.66 ± 0.13	0.60 ± 0.09	0.30 ± 0.11	0.51 ± 0.14	0.53 ± 0.14	0.28 ± 0.15
Random	Randomly pick between C HNR	0.62 ± 0.24	0.63 ± 0.18	0.26 ± 0.17	0.48 ± 0.28	0.48 ± 0.28	0.26 ± 0.18
Random	Randomly pick between C HWR	0.58 ± 0.14	0.55 ± 0.08	0.28 ± 0.12	0.61 ± 0.15	0.57 ± 0.12	0.32 ± 0.14