Ethical Implications of COMPAS Data Analysis

The two most ethically relevant findings of this project were (1) evidence of racial bias in COMPAS outputs and (2) the perfect independence of Display Text and Race, which may suggest that the data was constructed to produce this sort of independence. We found that the mean COMPAS score for non-white people was higher than that of white people (even when we subdivided the risk scores into those for violence, failure to appear, and recidivism), and that the probability of this result have arisen due to chance was almost 0. This is compelling evidence that the results of the algorithm are racially biased. This racial bias is likely a result of the algorithm's reflection of real-world bias in the US criminal justice system. The COMPAS algorithm was trained on real-world data that likely reflected a higher recidivism rate for non-white people, possibly because of racialized policing and other socio-economic disparities. This raises an ethical question that the statistical community is largely divided over: should algorithms correct for real-world bias or should they reflect them? The COMPAS algorithm is statistically sound, but is that the only criteria for success? Should the designers of the COMPAS algorithm have made their algorithm incorrectly model real-world data in order to correct for racial bias? Is the algorithm merely reflecting real-world bias or entrenching it?

Another ethical question arose because of the perfect independence between Display Text, which is what a judge would have seen as one output of the algorithm, and race. Our chi-square statistic was 0 and our p-value was 1. As we saw during Paul's R-scripts, constructing perfectly independent variables is almost impossible because even seemingly independent variables typically show some correlation in a chi-square test. The fact that the two variables in question are *perfectly* independent raises serious questions about the integrity of our dataset. It

seems nearly impossible that such perfect independence could have arisen due to chance: the data was most likely constructed in a manner that would reflect this independence. Either the designers of the algorithm created an algorithm free of any racial bias in Display Text (which seems unlikely given the racial bias present in the raw scores themselves), or whoever collected, cleaned, and released this data misconstrued the dataset to reflect this relationship (or lack thereof). This could be evidence of data fraud, which is an incredibly serious accusation to level. Our main ethical dilemma is due to the seriousness of the charge: if we have made a single mistake in our statistical tests that resulted in this independence, then we could potentially ruin someone's reputation and career for nothing. We would appreciate any advice on what to do with our findings, and any possible explanations for our results besides the worst one.