Transparency, Statistics, and Justice System Knowledge Is Essential for Science of Risk Assessment

Greg Ridgeway^{1,2}

¹Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America,

²Department of Criminology, College of Arts and Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Published on: Mar 31, 2020

DOI: <https://doi.org/10.1162/99608f92.cb0f8674>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

I first wish to congratulate Rudin, Wang, and Coker (2020, hereafter RWC) for a thorough analysis of COMPAS with the Broward County data. This is a much-needed analysis, as the ProPublica analysis (Angwin, Larson, Mattu, & Kirchner, 2016) has led to misconceptions about risk and needs assessments, while RWC bring into sharp focus the fundamental problem, lack of transparency.

RWC's article shows that advances in science come with transparency, the careful use of statistics as a part of data science, and a clear understanding of the application area (the justice system in this case).

The justice system is not transparent. Arrest decisions, bail assignments, pretrial detention decisions, plea bargain offers, and the sizes of fines and lengths of sentences all occur in an opaque process involving the police, prosecutors, magistrates, judges, and, of course, the defendants. Some of these decisions occur in court hearings with publicly accessible, complete transcripts, while others, such as plea negotiations, can occur by private text messages between a prosecutor and defense counsel. Each of the individuals working in the justice system may have clear principles in mind, but disparities and variability in justice system outcomes lead to the perception that the system is random or even unfair. Tools such as risk assessments and sentencing guidelines are intended to provide some standardization of justice system decisions.

The Pennsylvania sentencing guidelines suggest a prison term of at least 4 years for forcible rape and a prison term of at least 5 years for trafficking more than a kilogram of cocaine. We can argue whether the lengths of these terms are sensible, whether cocaine trafficking should have a more severe sentence than rape, and what aggravating circumstances should result in sentences above the minimums. Importantly, the sentencing guidelines are *transparent*. We all can see the relationship between the criminal offense and the recommended sentence.

Research has shown that sentencing guidelines change judges' behaviors. Bushway, Owens, and Piehl (2012) found that 10% of the worksheets used to extract sentencing recommendations for specific cases in Maryland had errors. These errors caused the calculated sentencing guidelines for prison sentences to be between one year too short to two years too long. Judges, unaware of these errors, imposed sentences that were shorter or longer depending on the direction of the errors. Data quality matters when producing decision-support tools, but it also shows that such tools are powerful influencers of judges. When public and political awareness of the sentencing disparity between crack and powder cocaine grew, legislation in Congress and executive action by the U.S. Sentencing Commission modified the sentencing guidelines in 2007, 2010, and 2018 and applied them retroactively in an effort to undue past disparities. Transparent tools such as sentencing guidelines and risk assessments can be used to encode our values, implement our evolving views on the criminalization of drug abuse, and push justice system reform.

As with the crack/powder cocaine disparity in sentencing guidelines now being undone, concerns abound that automated risk assessment tools are hardcoding historic (or even persisting) injustices into the justice system.

COMPAS is an obvious target since its use is widespread, it generates profits for its owners, and, as RWC point out, it is not transparent. However, using the current justice system as a benchmark, COMPAS does not make our justice system less transparent. Instead, COMPAS is actually a small step forward from the murky workings of the justice system. At least with COMPAS we know its complete set of inputs and can have this discussion about their relationship with its outputs.

Like the justice system, the lack of transparency makes COMPAS easy to misunderstand and for the journalists at ProPublica to assume that differences in error rates must mean that “it’s biased against blacks” (Angwin, Larson, Mattu, & Kirchner, 2016). It truly might be biased, but COMPAS would be more easily diagnosed if it was transparent. But ProPublica’s findings could be yet another case of Simpson’s paradox. RWC show convincingly that appropriately accounting for age leads to the conclusion that COMPAS “has at most weak dependence on race.” A simple numerical example can also show that observed disparities could be explained by confounding.

Table 1 shows ProPublica’s COMPAS error rates for White and Black defendants in Broward County. It shows that about 24% of ‘high-risk’ white defendants did not reoffend while 45% of ‘high-risk’ black defendants did not reoffend. A similar disparity holds for defendants that COMPAS rated ‘low risk.’

Table 1. COMPAS prediction fails differently for Black defendants.

	White	Black
Labeled Higher Risk, But Didn’t Reoffend	23.5%	44.9%
Labeled Lower Risk, Yet Did Reoffend	47.7%	28.0%

In Table 2, I have imagined four crime categories that have racial imbalances. Black defendants are more likely to be arrested and charged for crime categories 1 and 3, while White defendants are more likely to be charged for crimes 2 and 4. The risk of reoffending varies depending on the crime category, but, importantly, it does not vary by race. Regardless of race, defendants facing charges in crime categories 1 and 2 are considered low risk (less than 50% chance of reoffending) and those facing charges in crime categories 3 and 4 are considered high risk.

Table 2. Simulated example with race-neutral risk prediction that reproduces ProPublica’s error rates.

	Crime category
--	----------------

	#1	#2	#3	#4
Number of White defendants	103	2,021	855	2,021
Number of Black defendants	2,347	224	2347	83
Probability of reoffending, independent of race	0.26	0.49	0.54	0.86
Risk prediction, High when risk is greater than 50%	Low	Low	High	High

I have designed these reoffense probabilities to reproduce ProPublica's error rates. To see this, calculate for White and Black defendants separately; how many of them will be classified as high or low risk and how many will go on to reoffend? Table 3 shows the expected counts.

Table 3. Expected counts of cases classified by race, risk prediction, and future reoffense.

		High risk predicted	Low risk predicted
White	Reoffended	$855 \times 0.54 + 2,021 \times 0.86$ = 2,200	$103 \times 0.26 + 2,021 \times 0.49$ = 1,017
	Did not reoffend	$855 \times 0.46 + 2,021 \times 0.14$ = 676	$103 \times 0.74 + 2,021 \times 0.51$ = 1,107
Black	Reoffended	$2,347 \times 0.54 + 83 \times 0.86$ = 1,339	$2,347 \times 0.26 + 224 \times 0.49$ = 720
	Did not reoffend	$2,347 \times 0.46 + 83 \times 0.14$ = 1,091	$2,347 \times 0.74 + 224 \times 0.51$ = 1,851

Finally, Table 4 shows the expected error rates from this simulated example. Each error rate is nearly identical to the error rate ProPublica produced, but in this simulated example we know that the risk prediction is race neutral. It depends only on the crime category and the true risk of reoffense based on that crime category. This simple example involved only four crime categories when in reality there are many more. Also, the counts and

rates shown in Table 2 are not a unique set of numbers that reproduce ProPublica's error rates. This shows that racial disparities in error rates in the aggregate may be the result of confounding.

Table 4. Error rates observed from a race-neutral risk assessment.

	White	Black
Labeled Higher Risk, But Didn't Reoffend	676/(2,200+676) = 0.235	1,091/(1,339+1,091) = 0.449
Labeled Lower Risk, Yet Did Reoffend	1,017/(1,017+1,107) = 0.479	720/(720+1,851) = 0.280

To their credit, the data scientists at ProPublica made all of their data and analysis code available... completely transparent. This has allowed several teams in addition to RWC to probe their findings. Transparency has made it possible for the science of risk assessment to advance.

I take three key lessons from RWC's article.

First, transparency reduces the risk of misunderstanding. Unlike most statistical problems in which the unknown parameter can never really be fully known, COMPAS is knowable if only equivalent (formerly Northpointe) would reveal its structure. Only then can we be as sure over what we are arguing as we can with sentencing guidelines. ProPublica's linearity assumption likely led to an incorrect conclusion about racial bias, something completely avoidable if the COMPAS structure were known.

Second, RWC's analysis shows that statistics is an integral part of data science. As RWC unravel the ProPublica analysis in their article, it becomes clear that understanding COMPAS in Broward County required a world-class team of data scientists and statisticians in order to get closer to the truth. Data manipulation tools and a few basic modeling techniques resulted in ProPublica data scientists stumbling into Simpson's paradox without recognizing the risk.

Lastly, data science and statistics cannot ignore the need to have a clear understanding of the application, the justice system in this case. In some moments, RWC misunderstand part of COMPAS's purpose. COMPAS is a risk *and needs* assessment. Needs assessments are designed to help align treatment services with an offenders' key issues that likely lead to criminal behavior. RWC suggest that questions like "How hard is it for you to find a job ABOVE minimum wage compared to others?" should not be permitted. It would not be appropriate for assessing criminal risk, but it is an excellent question to understand whether a probation program with a strong job training component would be advisable. COMPAS also assesses history of victimization and abuse, financial instability, access to shelter and food assistance, substance abuse, and other needs. These are essential

for developing a comprehensive plan for offenders that go beyond sentencing decisions and should not be lost in the narrower discussion of risk assessment.

If we do not push forward with the kinds of thorough analyses demonstrated in this article and continue to advance risk prediction tools that are transparent and encode our justice system values, then, in the words of the authors, “we could be stuck in the situation where a human decision maker provides nontransparent, potentially biased decisions.”

Disclosure Statement

Greg Ridgeway has no financial or non-financial disclosures to share for this article.

References

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. Retrieved February 21, 2020, from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Bushway, S., Owens, E., & Piehl, A. (2012). Sentencing guidelines and judicial discretion: Quasi-experimental evidence from human calculation errors. *Journal of Empirical Legal Studies*, 9(2), 291–319. <https://doi.org/10.1111/j.1740-1461.2012.01254.x>
- Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.6ed64b30>
-

Addendum

April 29, 2020¹

ProPublica’s article labels their error rates like “Labeled Higher Risk, But Didn’t Reoffend.” Naturally, I interpreted the associated probabilities as $P(\text{Didn’t reoffend} \mid \text{Labeled Higher Risk})$, a false discovery rate. However, ProPublica separately produced a more detailed description of their analysis showing the underlying counts (Larson, Mattu, Kirchner, & Angwin, 2016). These details show that the probabilities in the original article were actually $P(\text{Labeled Higher Risk} \mid \text{Didn’t reoffend})$, a false positive rate. When asked in 2017 about the confusing labeling, the ProPublica author expressed no interest in a “semantic gotcha” (Larson, 2017) even though mistakenly reversing conditioning is a highly problematic issue in criminal justice reasoning, such as in the Prosecutor’s Fallacy (Thompson & Schumann, 1987).

My misinterpretation of ProPublica’s error rates means that the published numerical example in my original commentary is incorrect. The argument and its conclusions remain unchanged. That is, the error rates in the

ProPublica article might be entirely due to confounding. However, the numerical example that I used does not reflect the ProPublica rates correctly. Here I have generated a new numerical example that rectifies this.

Table 1 shows ProPublica's COMPAS error rates for White and Black defendants in Broward County. I have included both the ProPublica original labels as well as their probabilistic meaning confirmed from the raw data. It shows that about 24% of white defendants who did not reoffend were originally labeled as 'high risk' while 45% of black defendants who did not reoffend were originally labeled as 'high-risk.' A similar disparity holds for defendants who did reoffend.

Table 1. COMPAS prediction fails differently for Black defendants.

Original label	Actual meaning	White	Black
Labeled Higher Risk, But Didn't Reoffend	P(Labeled Higher Risk Didn't reoffend)	23.5%	44.9%
Labeled Lower Risk, Yet Did Reoffend	P(Labeled Lower Risk Did Reoffend)	47.7%	28.0%

In Table 2, I have imagined three crime categories that have racial imbalances. Black defendants are more likely to be arrested and charged for crime categories 2 and 3, while White defendants are more likely to be charged for crime 1. The risk of reoffending varies depending on the crime category, but, importantly, it does not vary by race. Regardless of race, defendants facing charges in crime category 1 are considered low risk (less than 50% chance of reoffending) and those facing charges in crime categories 2 and 3 are considered high risk.

Table 2. Simulated example with race-neutral risk prediction that reproduces ProPublica's error rates.

	Crime category		
	#1	#2	#3
Number of White defendants	3051	335	1614
Number of Black defendants	1959	1226	1816
Probability of reoffending, independent of race	0.42	0.63	0.74

Risk prediction, High when risk is greater than 50%	Low	High	High
--	-----	------	------

I have designed these reoffense probabilities to reproduce ProPublica's error rates. To see this, calculate for White and Black defendants separately; how many of them will be classified as high or low risk and how many will go on to reoffend? Table 3 shows the expected counts.

Table 3. Expected counts of cases classified by race, risk prediction, and future reoffense.

		High risk predicted	Low risk predicted
White	Reoffended	$335 \times 0.63 + 1614 \times 0.74$ = 1405	3051×0.42 = 1281
	Did not reoffend	$335 \times 0.37 + 1614 \times 0.26$ = 544	3051×0.58 = 1770
Black	Reoffended	$1226 \times 0.63 + 1816 \times 0.74$ = 2116	1959×0.42 = 823
	Did not reoffend	$1226 \times 0.37 + 1816 \times 0.26$ = 926	1959×0.58 = 1136

Finally, Table 4 shows the expected error rates from this simulated example. Each error rate is nearly identical to the error rate ProPublica produced, but in this simulated example we know that the risk prediction is race neutral. It depends only on the crime category and the true risk of reoffense based on that crime category. This simple example involved only three crime categories when in reality there are many more. Also, the counts and rates shown in Table 2 are not a unique set of numbers that reproduce ProPublica's error rates. This shows that racial disparities in error rates in the aggregate may be the result of confounding.

Table 4. Error rates observed from a race-neutral risk assessment.

Original label	Actual meaning	White	Black
Labeled Higher Risk, But Didn't Reoffend	P(Labeled Higher Risk Didn't reoffend)	$544 / (544 + 1770)$ = 0.235	$926 / (926 + 1136)$ = 0.449
Labeled Lower Risk, Yet Did Reoffend	P(Labeled Lower Risk Did Reoffend)	$1281 / (1405 + 1281)$ = 0.477	$823 / (2116 + 823)$ = 0.280

While there are large differences in false positive rates shown in Table 4, the false discovery rates, $P(\text{Didn't reoffend} \mid \text{Labeled Higher Risk})$, are nearly equal for black ($926/(926+2116) = 0.30$) and white ($544/(544+1405) = 0.28$) defendants. The underlying figures in the ProPublica article also show very little racial difference in false discovery rates (Larson, Mattu, Kirchner, & Angwin, 2016). It is well known that it is not possible to simultaneously achieve parity on common measures of fairness (Chouldechova, 2017; Kleinberg, Mullainathan, & Raghavan, 2017; Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017).

As I noted in my original discussion paper, the data scientists at ProPublica made all of their data and analysis code available. Had I probed more deeply into their original data and notes, I would have been able to see that the error rates reported were false positive and false negative rates.

Addendum References

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806). Association for Computing Machinery. <https://doi.org/10.1145/3097983.3098095>

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Dagstuhl, Germany: Schloss Dagstuhl--Leibniz-Zentrum fuer Informatik. <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>

Larson, J. (2017, March 17). On Twitter. Retrieved from <https://twitter.com/thejefflarson/status/842890560734183424>

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How we analyzed the COMPAS recidivism algorithm. *ProPublica*. Retrieved April 4, 2020, from <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

Thompson, W., & Schumann, E. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and defense attorney's fallacy. *Law and Human Behaviour*, 11(3), 167–187. <http://doi.org/10.1007/BF01044641>

©2020 Greg Ridgeway. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the

article.

Footnotes

1. Thanks to Dr. Sam Corbett-Davies for pointing out to me to my misinterpretation of the row labels in ProPublica's table. [↩](#)