PNAS COMMENTARY

Cohort bias in predictive risk assessments of future criminal justice system involvement

Jens Ludwig[a,1], Jon Kleinberg[b], and Sendhil Mullainathan[c]

In this issue of PNAS, a new paper by Montana et al. document how training a predictive risk tool in one time period and deploying it in another can lead to misprediction—"cohort bias." This can happen when the correlation between the outcome to be predicted (rearrest, etc.) and the predicator variables (characteristics of people, places, etc.) changes over time. The finding is important partly because of the magnitudes of misprediction Montana et al. document, and partly because of the growing use of these types of statistical models to inform decisions not just in criminal justice (who to release awaiting trial, who to let out on parole, who to prioritize for social programs) but in hiring, lending, school admissions, child protection, or medicine (1–5).

In principle, there should be a simple fix to this: Ensure that these statistical models are retrained regularly on new data as the world is changing over time. But, the very fact that there is such an easy fix raises a deeper question: Why is this fix, so simple in principle, ignored so often in practice?

The deeper problem that cohort bias helps illustrate are the deficiencies of our current system for evaluating algorithms with important policy consequences. We highlight three key deficiencies.

> **"Montana et al. document how training a predictive risk tool in one time period and deploying it in another can lead to misprediction—'cohort bias.'"**

The most important of these deficiencies is the lack of any systematic requirement that algorithms for policy intervention be evaluated before or after they are deployed. Consider how the US Food and Drug Administration requires new pharmaceuticals or medical devices to be tested with randomized controlled trials prior to deployment. That is complemented by other requirements for ongoing safety monitoring once products are deployed out in the field. A similar requirement for algorithms could reveal cohort bias as it unfolds over time. Since algorithms are at heart a policy intervention, our call for more rigorous and systematic algorithmic evaluation can be thought of as part of the broader push for more evidence-based policy (6, 7).

A second deficiency is that even in those cases where algorithms are evaluated, these evaluations are often limited by confusion between prediction quality and decision quality. Prediction quality is about how well the predictions of some algorithm relate to actual values, as measured by statistics such as calibration and area under the radio operating curve (AUC). But, algorithms deployed in many policy-relevant settings serve as decision aids for a human decision maker. For example, a judge is given a risk tool to help inform decisions about which defendants to release versus jail awaiting trial. The judge will inevitably follow the algorithm's recommendations sometimes but not always (8). Most relevant for policy is how the algorithm changes not prediction quality, but decision quality—for example, how does the tool change judge decisions? That can only be learned by evaluating algorithms as they are deployed, in situ (5, 9–12).

Finally, there are methodological issues that arise with the evaluation of algorithms specifically. For example, the outcome the algorithm is designed to predict may or may not match what enters into the policymaker's objective function, the problem of omitted payoff bias (13). Any attempt to evaluate an algorithm using retrospective data must account for the fact that the data themselves are generated by the past decisions of humans, who may have access to private information not captured in the dataset, which will selectively censor what outcome data are available—the selective labels problem. The full range of these evaluation challenges is not yet even fully understood, much less solved.

As many have argued, there are enormous opportunities for algorithms to positively impact social conditions (2). Humans are known to have trouble making predictions, are noisy in their decision-making, and also have implicit biases (14–16). The data show that algorithms have the potential to predict more accurately, and with less bias, than do humans (13, 17). But there is no guarantee any given algorithm will realize this potential (18).

If we do not demand stronger forms of evaluation for policy-relevant algorithms as they are deployed in practice, we create the risk of ineffective or even harmful algorithms in everyday use. It is therefore increasingly critical to not only use the right methodological tools to evaluate algorithms, but more generally to create appropriate incentives for designing and evaluating algorithms that come with stronger guarantees on their effects in use. With such systems in place, we expect that there can be many success stories for algorithms in addressing policy problems.

Author affiliations: [a]Harris School of Public Policy, University of Chicago and National Bureau of Economic Research (NBER), Chicago, IL 60637; [b]Department of Computer Science, Cornell University, Ithaca, NY 14853; and [c]Booth School of Business, University of Chicago and National Bureau of Economic Research (NBER), Chicago, IL 60637

Author contributions: J.L., J.K., and S.M. wrote the paper.

The authors declare no competing interest.

See companion article, "Cohort bias in predictive risk assessments of future criminal justice system involvement," 10.1073/pnas.2301990120.

[1]To whom correspondence may be addressed. Email: jludwig@uchicago.edu.

Published May 31, 2023.

1. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine bias" Pro Publica, posted May 23 (2016).
2. S. Athey, Beyond prediction: Using big data for policy problems. *Science* **355**, 483–485 (2017).
3. J. L. Koepke, D. G. Robinson, Danger ahead: Risk assessment and the future of bail reform. *Washington Law Rev.* **93**, 1725 (2018).
4. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
5. M.-P. Grimon, C. Mills, The impact of algorithmic tools on child protection: Evidence from a randomized controlled trial (Princeton University working paper, 2022).
6. J. M. Gueron, The politics and practice of social experiments: Seeds of a revolution (MDRC Working Paper, New York, 2016).
7. J. Baron, A brief history of evidence-based policy. *ANNALS Am. Acad. Polit. Soc. Sci.* **678**, 40–50 (2018).
8. J. Ludwig, S. Mullainathan, Fragile algorithms and fallible decision-makers: Lessons from the justice system. *J. Econ. Perspect.* **35**, 71–96 (2021).
9. A. Albright, If you give a judge a risk score: Evidence from Kentucky bail decisions. Harvard University Working Paper (2019).
10. V. Angelova, W. Dobbie, C. S. Yang, Algorithmic recommendations and human discretion (National Bureau of Economic Research Working Paper, Cambridge, MA, 2022).
11. M. Stevenson, Assessing risk assessment in action. *Minn. L. Rev.* **103**, 303 (2018).
12. M. T. Stevenson, J. L. Doleac, Algorithmic risk assessment in the hands of humans. Available at SSRN 3489440 (2022).
13. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, Sendhil Mullainathan, Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).
14. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).
15. J. Kleinberg, J. Ludwig, S. Mullainathan, C. R. Sunstein, Discrimination in the age of algorithms. *J. Legal Anal.* **10**, 113–174 (2018).
16. D. Kahneman, O. Sibony, C. R. Sunstein, *Noise: A Flaw in Human Judgment* (Little, Brown Spark, 2021).
17. D. Li, L. R. Raymond, P. Bergman, Hiring as exploration (National Bureau of Economic Research Working Paper, Cambridge, MA, 2020), p. 27736.
18. E. Montana, D. S. Nagin, R. Neil, R. J. Sampson, Cohort bias in predictive risk assessments of future criminal justice system involvement. *Proc. Natl. Acad. Sci. U.S.A.* (2023).