Reply to the discussants of "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment."

Journal of the Royal Statistical Society, Series A, **Forthcoming.**

Kosuke Imai* Zhichao Jiang† D. James Greiner‡ Ryan Halen§ Sooahn Shin¶

February 17, 2023

We thank the Research Section of the Royal Statistical Society for providing us with a valuable opportunity to present our paper and receive feedback from many scholars. Despite their diverse perspectives, we believe that all of our discussants are in agreement about the need to develop new statistical methodology to better analyze algorithm-assisted human decision making.

While the use of algorithms has become ubiquitous in today's society, we — humans — still make many consequential decisions with the help of algorithms. As Kumar and VanderWeele noted, such a hybrid decision-making system may allows one to combine human experiences and knowledge with algorithmic recommendations, possibly leading to improved decisions. Moreover, even if fully algorithmic decision-making is more optimal than a hybrid system, we may still prefer the latter because we want to hold humans, rather than algorithms, accountable for the consequences of decisions.

The importance of algorithm-assisted human decision making motivated us to develop a set of methodological tools to evaluate and understand how algorithmic recommendations affect human decisions. As echoed by Cruz Cortéz and Gosh, our study sharply contrasts with much of the existing studies whose focus has been the accuracy and fairness of algorithms themselves. We hope other researchers follow up on this important research agenda. Below, we respond to specific comments raised by the discussants.

Scientific and policy-relevant questions. When analyzing algorithm-assisted human decision making, one important question is whether algorithmic recommendations help humans make better decisions. In the context of the pretrial public safety assessment (PSA), the goal of judge's decision is to prevent crimes while avoiding unnecessarily harsh decisions. If the PSA helps them achieve this goal, it should nudge judges to make a harsher decision in the case that would prevent an arrestee from committing a crime while encouraging them to make a more lenient decision for those who would not cause a negative outcome regardless of judge's decision. Our causal estimands

*Corresponding author. Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu URL: https://imai.fas.harvard.edu

†Professor, School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong 510275, China. Email: jiangzhch7@mail.sysu.edu.cn

‡Honorable S. William Green Professor of Public Law, Harvard Law School, 1525 Massachusetts Avenue, Griswold 504, Cambridge, MA 02138.

§Data Analyst, Access to Justice Lab at Harvard Law School, 1607 Massachusetts Avenue, Third Floor, Cambridge, MA 02138.

¶Ph.D. student, Department of Government, Harvard University, Cambridge, MA 02138. Email: sooahnshin@g.harvard.edu URL: http://sooahnshin.com

(Average Principal Causal Effects or APCEs) directly addresses this key question of whether and how algorithmic recommendations help human decision makers achieve their goal. Our methodology shows how to learn about these causal quantities from the observed data.

Some discussants such as Dawid argue that the simple comparison of human decisions (and the downstream outcomes) with and without algorithmic recommendations can answer this question. In the case of our randomized controlled trial, they suggest that we estimate the causal effect of PSA provision on decisions by judges and the negative outcomes (see Figure 2 of our article for the results of this analysis). Unfortunately, although such an intention-to-treat (ITT) analysis evaluates whether the provision of algorithmic recommendation affects judicial decisions and/or arrestees behavior, it cannot answer the essential question of whether PSA provision helps judges achieve their goal. Suppose, for example, that the ITT analysis shows the provision of PSA leads to harsher decisions and reduces the number of crimes. This evidence alone cannot rule out the possibility that the PSA provision leads to many unnecessarily harsh decisions. To properly evaluate the efficacy of PSA, we must understand whether it encourages judges to make a harsh decision only when it helps prevent a crime. This is exactly what our proposed methodology accomplishes.

More generally, as many discussants (e.g., Cuellar, Cruz Cortéz and Gosh, Ding, Hunsicke, Kumar, Python, VanderWeele, and Wijayatunga) recognize, the evaluation of algorithm-assisted human decision making requires analyzing the *mechanisms*, through which algorithmic recommendations influence human decisions. Such an analysis is critical for policy makers who are considering the adoption of algorithmic recommendations. It is also essential for social scientists whose goal is to understand how humans incorporate algorithmic recommendations into their decisions. Simply estimating the overall ITT effects of algorithmic recommendation on human decisions and downstream outcomes is not sufficient for these purposes.

Essential role of potential outcomes and principal stratification. Some discussants dispute the value of our methodology and advocate an alternative approach. Dawid criticizes the use of principal stratification while noting that "some of the proponents of these arguments have been honoured with the 2021 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel." We let readers decide whether approaches based on potential outcomes and principal stratification, including ours, are useful or "totally misguided." Instead of engaging with this philosophical debate, we examine Dawid's recommendation to use the statistical decision theory for answering the following questions: (1) "should we present PSA to the judge before she makes her decision, or not?", and (2) "How, ideally, should the judge use the information she has in coming to her decision?"

Dawid is incorrect to assume that his first question can be answered by the ITT analysis alone. As explained above, to decide on the use of PSA, policy makers need to understand if PSA helps judges achieve their goal of preventing a negative outcome while avoiding unnecessarily harsh decisions. The ITT analysis alone cannot answer this key question of the PSA evaluation study.

Answering Dawid's second question also requires going beyond the analysis he suggests. As noted earlier, although a harsher decision typically reduces the probability of a negative outcome, a judge does not wish to issue the harshest decision to every arrestee. In fact, one goal of the PSA is to help a judge identify arrestees who would commit a crime only if they are given a lenient decision. According to this goal, an optimal decision needs to balance the cost of mistakenly detaining an arrestee who would not commit a crime regardless of the decision, against the cost of releasing an arrestee who would commit a crime only when released. Thus, joint potential outcomes are necessary to characterize the risk levels of arrestees.

In addition, theses costs may be different. For example, the cost of mistakenly detaining "safe" arrestees may exceed that of releasing "risky" arrestees. In medicine, the Hippocratic principle to "do no harm" has the same asymmetric utility structure: mistakenly killing a person is considered more costly than failing to save an individual. In Ben-Michael, Imai and Jiang (2022), we consider the optimal decision rule under such settings. Although the observed data do not identify the joint distribution of potential outcomes, we propose first to partially identify the expected utility and then

to find the optimal policy in the worst scenario. Any thoughtful evaluation of algorithm-assisted human decision making must directly address whether algorithmic recommendations help achieve the goal of human decision makers. Potential outcomes and principal stratification are useful tools for conducting such evaluation under various settings.

We briefly respond to Stensrud, Didelez, and Sarvet who make a similar criticism about the use of joint potential outcomes. They propose alternative "single-world" estimands, including $\{Y_i(Z_i = z) = D_i(Z_i = z) = 1\}$ which they refer to as "failed detention." But, in what sense, does this event represent a failure? If an arrestee is someone who would commit a crime regardless of the decision, it is unclear if the decision to detain this person is a failure. Without consideration of counterfactual outcome, one cannot determine whether a correct decision is made for each case.

Stensrud et al. further suggest that analysts focus on the same decision problems as the ones proposed by Dawid. They show that there exist direct relations between these decision problems and our causal quantities of interests. Although the existence of such relations is interesting, this does not alter the fact that without considering joint potential outcomes, one cannot even define causal quantities of interest related to the goal of judges — prevent negative outcomes while avoiding unnecessarily harsh decisions. More importantly, these relations exist under our identification strategies. In many causal analyses, it is useful to first define causal quantities of interest, using potential outcomes, and then consider how various assumptions can partially identify or point-identify them.

We disagree with Dawid, Didelez, Sarvet, and Stensrud who object to the use of principal stratification in general. In contrast, we believe that principal stratification and potential outcomes are a powerful language for causal inference. Many researchers studied causal quantities of interest that involve both potential outcomes, including individual treatment effects $(Y_i(1) - Y_i(0))$ (e.g., Lei and Candes, 2021), the proportion of those who harmed by the treatment $(\Pr(Y_i(1) - Y_i(0) < 0))$ (e.g., Kallus, 2022), quantiles of treatment effects (e.g., Fan and Park, 2012), and probability of necessary and sufficient causes $(\Pr(Y_i(t) = y \mid Y_i(t') = y'))$ (e.g., Pearl, 2000) to name a few.

As these examples and their scientific applications demonstrate, consideration of joint potential outcomes is an essential aspect of causal inference. Instead of engaging in philosophical debates, we strongly encourage these discussants to get directly involved in applied causal inference research through direct collaboration with policy makers and scientists. We believe that doing so will help them gain a deeper appreciation of the important roles played by potential outcomes and principal stratification in determining causal estimands and communicating scientific and policy-relevant findings.

Verifiability of identification assumptions. In response to the critiques by Stensrud, Didelez, and Sarvet, who highlight the unverifiability of our identification assumptions, we emphasize that both randomization of treatment assignment (Assumption 1) and exclusion restriction (Assumption 2) are guaranteed to satisfied under the study design. The monotonicity assumption (Assumption 3) is based on the domain knowledge (It is worth noting that many researchers in this literature assumes a stronger monotonicity condition that an arrestee cannot commit a crime if detained (e.g., Coston et al., 2020)). Under these three assumptions, we have shown that the sign of our causal quantity of interest is identified. Ding, in his discussion, further shows how to derive sharp partial identification bounds in more general settings.

Finally, unconfoundedness (Assumption 4) is also guaranteed to hold if analysts have access to all the information used by a judge when making the decision. We are currently in the process of collecting additional information including the transcripts of court proceedings. In our paper, we develop a sensitivity analysis that allows one to examine the robustness of empirical findings to possible violation of this assumption. In causal inference, partial identification bounds and sensitivity analysis play a crucial role in examining the significance of different identification assumptions.

Fairness and ethics. Some discussants raise the issues about fairness. In our paper, we used the concept of principal fairness, which was introduced in a separate paper (Imai and Jiang, 2022). The paper is forthcoming in *Statistical Science*. Cuellar points out a large literature on fairness in

computer science and statistics and wonders why we used principal fairness. We briefly respond here while leaving a more detailed discussion to the aforementioned paper. The main advantage of principal fairness is that it directly takes into account how the decision affects individuals. This consideration is critical in our study because judges' decisions can affect individual arrestees. In contrast, most existing fairness definitions, which are based on observed, rather than potential, outcomes, are unable to incorporate the causal effect of decision on the outcome of interest.

In addition, Huffton suggests that when considering fairness, risks associated with men and women should be considered separately. Stensrud, Didelez, and Sarvet point out that while an earlier version of our *Statistical Science* paper considered conditional principal fairness, our analysis focuses on the marginal principal fairness. The question of what variables to condition on is not statistical. Rather, it is an important question that confronts policy makers and the society. Finally, Jose suggests that fairness should be incorporated either in the utility or in the constraint when considering optimal decisions. We have written a related paper that incorporates principal strata in policy learning (Ben-Michael, Imai and Jiang, 2022). This framework can directly incorporate the principal fairness criteria into policy learning.

King raises ethical objections to our randomized controlled trial (RCT), expressing skepticism about Institutional Review Board (IRB) review. He suggests that because our study did not test alternative ways of providing the risk assessment instrument information to the judges (VanderWeele discusses an alternative presentation scheme as well), it is futile and therefore ethically objectionable. King also expresses concerns that the consequences at issue in our study (we presume King refers to incarceration) are too high to allow for randomization. Finally, he hints at problems with informed consent.

Harvard University's IRB conducted a thorough review of this study before authorizing it. They found our study to be justified based on many of the grounds briefly summarized below. Concerns of practicality and statistical power prevented us from testing alternate information presentation schemes. Regarding the remainder of King's concerns, we refer readers to Lynch, Greiner and Cohen (2020) for a more complete discussion, which anticipate and respond to them. In brief, a well-developed literature in medical ethics provides principles that can be applied to the legal setting. For example, equipoise is one of the oldest and most accepted principles, upon which ethicists and practitioners rely to support clinical trials of new drugs and medical devices. This principle states that when experts and field operators lack knowledge about whether an intervention will improve outcomes, randomization of the intervention is ethically permissible. The reason is that if the effectiveness of an intervention is unknown, then there is no relevant sense in which a practitioner or a researcher knowingly harms an individual by either applying or withholding the intervention. No one, including us, likes the fact that we frequently are in equipoise as to the effectiveness of interventions, but the way to address this unfortunate situation is with credible research designs, not ethically induced paralysis.

Principles of medical research ethics also answer King's concern that the seriousness of an outcome to be measured in a study (here, incarceration) renders randomization impermissible. Assuming that additional incarceration is undesirable (this is disputed, at least in the United States, and its resolution requires a complex balancing of interests), the argument is coherent but proves too much, perhaps even for its proponents. A typical outcome in a cancer drug study is death within a certain number of years. If incarceration is too consequential an outcome to allow a randomized design, then death is as well. The fact that the consequences of an intervention are high is a reason to implement credible research designs, not to avoid them. The use of RCTs, even in cases where the consequences are high, is crucial for rigorous evaluation of interventions.

We did not seek to elicit informed consent from the arrested individuals subject to the judge's decision for several reasons. First, providing the disclosure needed to make a consent process informed was not practicable in the context of a three-minute criminal first appearance hearing where the primary interest of the arrested individual is to persuade the judge to release them from custody. Second, the intervention in this study is on the judge, not the arrested individual, and thus is it is

not clear that the arrested individual is a participant in the study as opposed to a bystander whose interests may be affected in a similar way as, for example, a person who lives in close proximity to a participant in a vaccine/infectious disease RCT. Third, unfortunately, the criminal justice setting is by social design highly coercive. Within very wide legal limits, the arrestee has no right or legitimate expectation that a judge will or will not have access to particular information when making an initial release decision, and randomization of the information provided to a judge in an arrestee's case deprives the latter of nothing.

Finally, several commentators asked why our study design did not involve blinding of judges either to the intervention or to the fact of the study. First, it was not possible to keep the judges unaware of the presence of the risk assessment instrument in some cases but not others. Second, according to ethical principles, if an individual's decision-making is the focus of a study, researchers should, whenever possible, obtain their consent to participate in the study. In this context, eliciting informed consent from the judges requires informing them about the study.

Future research directions. There are several avenues for future research. The first is the question of how to improve the way in which the PSA is calculated. The key challenge lies in the fact that like many algorithmic recommendations used in public policy, the PSA is a deterministic rule. This means that learning a better rule from the observed data requires some degree of extrapolation. In Ben-Michael et al. (2021), we use the minimax statistical decision theory to limit the probability that the learned new rule performs worse than the existing rule. This is achieved by first specifying a model class for the conditional average treatment effect, which partially identifies the counterfactual outcome, and then maximizing the expected utility in the worst case using robust optimization. Another possibility is to present the algorithmic recommendations to judges in a different way. For example, VanderWeele suggests that the use of predicted probabilities, rather than integer values as done in the PSA, might better convey the relevant information to human decision makers.

Second, the proposed methodology can be extended to reflect additional real world complexities. For example, judges make decisions about monitoring conditions as well as bail amount. We can also analyze multiple outcomes jointly rather than one outcome at a time as done in the current paper. Such generalization of our methodological framework will allow us to provide a more holistic evaluation of the PSA. For example, Ding generalizes our partial identification result by showing how to compute sharp bounds under the ordinal decision without additional assumptions. One may apply a similar analytical strategy to even more general settings such as multidimensional decisions and outcomes.

Third, although we have focused on the analysis of first arrests in our article, a relatively large number of people are rearrested in our data. Developing statistical methods for dealing with multiple arrests is important because it enables the use of all available data and answers important questions related to recidivism and accumulated impact of PSA on judges' decisions and arrestees' behavior.

Finally, Conti raises an important issue about the external validity while Cuellar emphasizes the value of conducting additional experiments with more judges and jurisdictions. In particular, they question how generalizable our empirical findings are given that the data come from Dane county, Wisconsin alone. We are currently conducting multiple experiments in other jurisdictions and hope to tackle the challenge of developing new methods to learn from multiple experiments in different sites.

References

Ben-Michael, Eli, D. James Greiner, Kosuke Imai and Zhichao Jiang. 2021. Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment. Technical Report. arXiv:2109.11679.

Ben-Michael, Eli, Kosuke Imai and Zhichao Jiang. 2022. Policy learning with asymmetric utilities. Technical Report. arXiv:2206.10479.

Coston, Amanda, Alan Mishler, Edward H Kennedy and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. pp. 582–593.

Fan, Yanqin and Soo Park. 2012. "Confidence intervals for the quantile of treatment effects in randomized experiments." *Journal of Econometrics* 167:330–344.

Imai, Kosuke and Zhichao Jiang. 2022. "Principal Fairness for Human and Algorithmic Decision-Making." *Statistical Science*.

Kallus, Nathan. 2022. What's the Harm? Sharp Bounds on the Fraction Negatively Affected by Treatment. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Lei, Lihua and Emmanuel J. Candes. 2021. "Conformal inference of counterfactuals and individual treatment effects." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 83:911–938.

Lynch, H. Fernandez, D. J. Greiner and I. G. Cohen. 2020. "Overcoming obstacles to experiments in legal practice." *Science* 367:1078–1080.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.