Expert Report

Nandita Mitra, PhD and Jason Roy, PhD

December 8, 2023

1 Summary Statement of Qualifications

Nandita Mitra, PhD

I am a Professor of Biostatistics in the Department of Biostatistics, Epidemiology, and Informatics at the Perelman School of Medicine at the University of Pennsylvania (Penn). I have a secondary faculty appointment in the Department of Statistics and Data Science at the Wharton School at Penn. I am also Co-Director of the Center for Causal Inference and was formerly Vice Chair of Education, Vice Chair of Faculty, and Chair of the Graduate Group in Epidemiology and Biostatistics at Penn.

I received an A.B. in Mathematics from Brown University, an M.S. in Biostatistics from the University of California, Berkeley, and my Ph.D. in Biostatistics from Columbia University. I also completed a post-doctoral fellowship at Harvard.

My primary research interests include the design and analysis of observational studies, causal inference, and statistical approaches for cost-effectiveness analysis. I have developed doubly robust approaches to estimation of cost-effectiveness measures, nonparametric influence function based instrumental variable estimators for censored outcomes, model-based sensitivity analysis approaches, and semi-parametric methods for assessing policy interventions under interference. I collaborate with investigators in oncology, health policy, and health economics and have co-authored over 300 peer-reviewed publications.

I am a leader in several international statistical organizations including: Chair of the Budget & Finance Committee of the International Biometrics Society, Chair of the American Statistical Association Statistics in Epidemiology Section, and Secretary of the Society for Causal Inference. In addition, I am the Editor-in-Chief of the journal *Observational Studies*. I was elected Fellow of the American Statistical Association (ASA) in 2019. The designation of ASA

Fellow has been a significant honor for nearly 100 years; the ASA can only elect up to one-third of one percent of the total association membership as fellows each year.

I have not served as an expert witness in the past.

Jason Roy, PhD

I am a Professor of Biostatistics and Chair of the Department of Biostatistics and Epidemiology at Rutgers University. I was formerly a professor of Biostatistics at the University of Pennsylvania where I was the founding Director of the Center for Causal Inference. I received my PhD in Biostatistics from the University of Michigan in 2000.

My primary research interests include causal inference methods for observational studies, missing data methods, and machine learning. I collaborate with researchers in the health and social sciences and have over 150 peer-reviewed publications. I co-authored a book in 2023 "Bayesian Nonparametric Methods for Causal Inference and Missing Data."

I developed the first online course on causal inference, "A Crash Course in Causality," that has had enrollment of over 38,000 people. Based on that course, I won the Causality in Statistics Education Award from the American

Statistical Association in 2021. I was elected Fellow of the ASA in 2019. I am Associate Editor for Biometrics and on the Executive Editorial Board for Observational Studies.

I have not served as an expert witness in the past.

2 References

1. Unnever, J. D., & Cullen, F. T. (2007). Reassessing the Racial Divide in Support for Capital Punishment: The Continuing Significance of Race. Journal of Research in Crime and Delinquency, 44(1), 124–158.

2. White, I.R., Royston, P. and Wood, A.M. (2011), Multiple imputation using chained equations: Issues and guidance for practice. Statist. Med., 30: 377-399.

3. Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika. 70(1):41-55.

4. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. Stat Appl Genet Mol Biol. 2007; 6: Article 25. doi: 10.2202/1544-6115.1309.

5. Arvid Sjölander, Sander Greenland, Are E-values too optimistic or too pessimistic? Both and neither! International Journal of Epidemiology, Volume 51, Issue 2, April 2022, Pages 355–363.

6. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. Epidemiology 2016;7:368–77.

7. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 2017;167: 268–74.

8. Ioannidis J, Tan Y, Blum M. Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. Ann Intern Med 2019;170:108–11.

9. Greenland S. Commentary: An argument against E-values for assessing the plausibility that an association could be explained away by residual confounding. Int J Epidemiol 2020;49:1501–03.

10. MacLehose RF, Ahern TP, Lash TL, Poole C, Greenland S. The importance of making assumptions in bias analysis. Epidemiology 2021;32:617–24.

11. Morgan, P. L. (2021). Unmeasured Confounding and Racial or Ethnic

Disparities in Disability Identification. Educational Evaluation and Policy Analysis, 43(2), 351-361.

12. Li F., Li F. (2023). Using propensity scores for racial disparities analysis. Observational Studies 9(1), 59-68.

3 Analysis

Introduction

We were asked to review the expert report and analyses of Dr. Fan Li. Dr. Li's report addressed a study of race and jury selection by researchers at the Michigan State University College of Law (MSU). Specifically, we reviewed the report written by MSU and analyzed data provided by MSU on November 27, 2023.[1] We also were provided and reviewed (i) a 2012 Report by Christopher Cronin, Ph.D. entitled "Black American Political Ideology, Partisanship, and Justice", (ii) statements and affidavits by prosecutors, and (iii) an excel spreadsheet compilation of the reasons given by prosecutors for striking Black

[1] We first analyzed the MSU dataset dated September 2023. We reference the results from the November dataset for this report but we have disclosed statistical files from our code with the September dataset as well as the November dataset.

jurors based upon the prosecutor statements.

Here we provide a summary of our analysis of the data, describe our methodology, and provide our conclusions in the context of Dr. Li's report. All R code and results can be found in the appendix (R Markdown document).

Association versus Causation in Racial Evaluations

Our objective in this analysis was to show whether or not Black venire members were more likely to be struck for reasons that cannot be explained by factors unrelated to race. For instance, if Black venire members happen to be older, on average, than non-Black members, we would want to make sure that differences in strike rates between racial groups is not simply due to their age differences. In any analysis, we would want to account for differences in these types of variables. On the other hand, we would *not* want to control for variables that are directly related to racial discrimination. For instance, there is significant literature documenting the fact that attitudes toward the death penalty differ significantly among racial groups as the result, in part, of the historical experiences of discrimination. (Cronin Report and testimony, 2012; Unnever & Cullen, 2007). Accordingly, it is not appropriate to control for death penalty reservations as a factor unrelated to race. In addition, it is inappropriate to control for biological or cultural factors that could be related to race, such as appearance or body language. In that sense, we disagree with Dr. Li's primary concern with the inability of the jury study to control for "venire member's physical appearance, manner, and body language." Li, p. 12. Physical appearance, manner, and body language are strongly tied in with racial perceptions and discrimination.

In a recently published paper, Dr. Li (Li and Li, 2023) notes that "Because race is not manipulable, racial disparity investigations are inherently not causal." She goes on to propose propensity score methods in order to balance factors between racial groups to be able to estimate the association between race and outcomes. Contradictorily, in her expert witness report, she states "I conclude that it is crucial to focus on the causal instead of associational role in determining the role race plays on jury selection." Li, p.12.

Because of the inherent difficulty and controversy in thinking about race in the counterfactual context, in this analysis we focus on robustly estimating the *association* between race and strike decision, using propensity score weighting to achieve balance on all important measured factors. We then argue that it is highly unlikely that there are other non race-related factors that we did not account for that would explain away the strong association.

Measures of Association

With a binary factor (venire member is Black or not Black) and binary outcome (venire member is struck or not), there are two common measures of association, the odds ratio (OR) and relative risk (RR). Our data can be presented as:

	Struck	Not struck
Black	a	b
Not Black	c	d

The RR is defined as the ratio of risks or probabilities:

$$RR = \frac{\textit{Probability of being struck if Black}}{\textit{Probability of being struck if not Black}} = \frac{a/(a+b)}{c/(c+d)}$$

On the other hand, the OR is defined as a ratio of two odds:

$$OR = \frac{\textit{Odds of being struck if Black}}{\textit{Odds of being struck if not Black}}$$

$$= \frac{\textit{Number of Black members struck} / \textit{Number of Black members not struck}}{\textit{Number of nonBlack members struck} / \textit{Number of nonBlack members not struck}}$$

$$= \frac{a/b}{c/d}$$

We note that Dr. Li presents a RR rather than the OR which is presented by Grosso and O'Brien. Dr. Li concludes that her smaller RR (1.45) versus Grosso and O'Brien's larger OR (2.48) is explained by her use of better causal methodology and better adjustment for confounders. Li, pp 19-20. However, her conclusion is not fully justified. It is well understood that when the outcome (peremptory strike) is common (greater than 10%), the OR will be larger than the RR. Hence, the discrepancy in values may be partially attributed to the measure that is presented (OR versus RR) rather than the under-adjustment of confounders.

In fact, when we conducted a standard multivariable regression analysis mimicking Grosso and O'Brien using the November data, we estimate an OR = 2.79 (2.27, 3.44)[2] and RR = 1.62 (1.46, 1.79).

Even though the OR measure is much larger than the RR, it is important to note that both measures demonstrate that there is a much higher and statistically significant probability (or odds) that Black venire members are struck versus non-Black members.

We elect to present the estimated RR to be consistent with Dr. Li's report.

[2] The 95% confidence intervals are reported in this report parenthetically.

Missing Data Imputation

We first evaluated the magnitude and patterns of missing data. We used a common and well proven approach using chained equations to conduct conditional mean imputation and predictive mean matching (White et al, 2011). This approach ensures that that the uncertainty in the imputed values is accounted for in the analysis and ensures that the imputed values are plausible given the observed data. Dr. Li imputed missing values by using the mean of the observed values for continuous variables and randomly drawing values for categorical variables. Our missing data imputation approach gave very similar results.

Propensity Score Analysis

We use propensity score weighting to balance differences between Black and non-Black venire members to be able to better compare these two groups.

In our analysis, we use an approach for estimating the propensity score that uses an ensemble machine learning approach called Super Learner (van der Laan, 2007). Super Learner combines multiple a priori-specified multivariable predictive algorithms into one optimized algorithm that returns a prediction function with the best cross-validated performance. It allows much more flexibility than standard approaches, such as logistic regression, to estimate the propensity score and allows the inclusion of interaction terms and higher-order terms to obtain the best estimate of the propensity score. In our analysis, we use generalized linear models, random forests, LASSO, and elastic net learners. Super Learner uses cross-validation to fit the models and eliminates judgment or preferences of the researcher in determining which model is best.[3] We then created both average treatment effect on the treated (ATT) and Overlap weights. There is good justification for both types of weights and both weights seem to perform very similarly in achieving balance and overlap. We then calculated standardized mean differences (SMDs) under both weights. All 28 variables achieved balance after weighting with all SMDs having values well below the standard rule of thumb of SMD < 0.1. This analysis using Super Learner and Overlap weights resulted in a RR of 1.58 (1.46, 1.70). In other words, after rigorously accounting for the 28 factors that could affect the association, we find that **Black venire members were 58% more likely to be struck than non-Black members.**

Both MSU and Dr. Li treated death penalty reservations as a factor to be

[3]We observed that Dr. Li ran more analyses than she included in her report and appeared to weigh her preferred result when selecting models based upon the desired outcome.

adjusted in their analyses. To be consistent with this approach and to compare the results of our independent propensity score analysis, we initially included death penalty reservations in the analysis. However, research shows large differences in death penalty reservations across racial groups, tied to historical and structural racism. (Cronin, 2012; Unnever & Cullen[4], 2007) Accordingly, we do not believe it is appropriate for an investigation of the role of race to treat death penalty reservations as a possible race neutral confounder. When we remove death penalty reservations from the list of variables that we control for, the RR increases to 1.75 (1.62, 1.89). In other words, **Black venire members were 75% more likely to be struck than non-Black members**.

[4]Unnever & Cullen state "Similar to previous research (Bobo and Johnson, 2004; Bohm 1987; Bohm, Clark, and Aveni 1991; Roberts and Stalans 1997), we show that there is a racial divide in support for the death penalty and that this divide is sustained even when a range of factors are controlled (see also Cochran and Chamlin 2006). Most important, the investigation reveals that factors that should lessen the racial divide, such as African Americans and Whites having the same religious affiliation, do not equally affect their attitudes, and indeed these potential ameliorating factors may result in further polarizing White and Black opinions toward the death penalty. Taken together, these findings suggest not only that a racial divide exists but also that is it unlikely to narrow even if African Americans and Whites converge in social position and social values. Indeed, it appears that African Americans and Whites have widely different collective biographies that uniquely influence their opinions of the death penalty (Hunt 1996; Smith and Seltzer 2000). Specifically, we propose that a key factor in sustaining the cleavage in capital punishment attitudes is the historical legacy of racial oppression that prompts African Americans in diverse social and cultural locations to be wary of the state's use of lethal punishment."

Across each of the approaches – the logistic regression favored by MSU or the propensity weighting that we performed and Dr. Li performed – race remained a robust predictor of strike decisions as reflected in Table 1.

Table 1: Relative Risks of Race on Strike Decisions

Controlling for Death Penalty Reservations as well as other factors			Controlling for other factors but not death penalty reservations
Grosso & O'Brien – Relative Risk (logistic regression)	Li Relative Risk (propensity score)	Mitra Roy Relative Risk (propensity score)	Mitra Roy Relative Risk (propensity score)
1.62	1.45	1.58	1.75

Results Stratified by Whether the Defendant was Black

To further explore the relationship between race and jury selection, we conducted an analysis to compare the effect of Black race on peremptory strike decisions in cases where the defendant was Black and where the defendant was not Black. To do this, we carried out separate analyses for cases with Black defendants and non-Black defendants. These analyses were conducting excluding death penalty reservations as discussed above. In cases where the defendant was not Black, the RR was 1.58 (1.42, 1.78). When the defendant was Black, the RR increased to 1.94 (1.75, 2.16). In other words, **Black venire members were 94% more likely to be struck than non-Black members when the defendant was Black**.

Unobserved Confounding and The E-value

Dr. Li claims that the findings of Grosso and O'Brien are spurious because there could be unobserved confounding (i.e., an unobserved factor that both influences race and the outcome). Dr. Li lists "venire member's appearance, manner, and body language" as factors that cannot be accounted for. Li, p 12. However, these factors can be thought of as proxies for racial discrimination and may be highly correlated with how race is perceived. Further, in Dr. Li's causal diagram (p. 8), the arrow directed from the unobserved confounder (e.g. venire member's appearance or manner) to race is not justifiable. Someone's appearance does not "cause" their race (Figure 1). If anything, the unobserved factors that she lists are potentially downstream from race and would not be considered to be confounders (Figure 2).

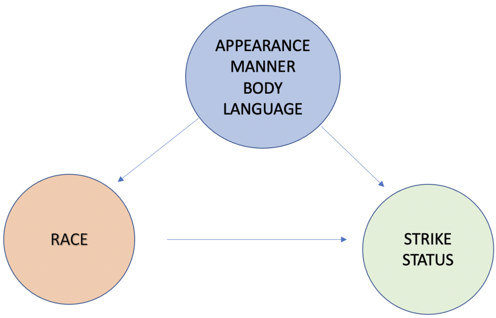

Figure 1: Confounding

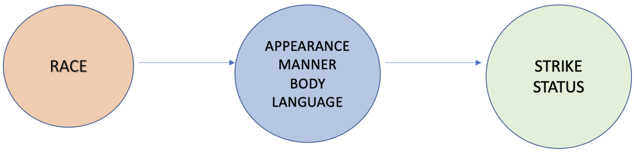

Figure 2: Not Confounding

Despite an unconvincing argument that there could be substantial unobserved confounding, Dr. Li uses the E-value to demonstrate the potential effect of unobserved confounding on her findings. A large E-value implies that considerable unobserved confounding would be needed to explain away the magnitude of the previously estimated RR.

Dr. Li computes the E-value to be 2.3 based on the findings of Grosso and O'Brien. Li, p 22. We have calculated the E-value using our estimated RR and found it to be 2.9. This means that the effect of race on strike decision would only be explained away by an unobserved confounder that is associated with both race and peremptory strike by a risk ratio of 2.9-fold each (with a lower bound of 2.6), after adjustment for observed confounders. Dr. Li goes on to say that "the existence of an unobserved confounder between race and jury selection of the strength in terms of E-value of 2.3 is highly plausible". Li, p 22. We strongly disagree with this statement. An E-value of 2.3, after accounting for all measured confounders (we accounted for 25 such confounders) is an extremely strong and highly unlikely association with a hypothetical unobserved confounder.

Greenland (2020) argues that the E-value may give an "unnecessarily pessimistic impression of the study". He argues that the effect of an unobserved confounder can be much weaker than proposed *if it is strongly correlated with measured confounders* that are accounted for in the analysis. MacLehose et al (2021) state: "The calculation of E-values for known but unmeasured confounders is irresponsible, as it makes no use of the information on those covariates that make them plausible to view as confounders. A desire for sensitivity analyses without assumptions is a desire to do inference in basic ignorance of background context".

If, as Dr. Li, claims, that venire members' appearance, manner, and body language are important unobserved confounders, it can be easily argued that these are factors that are highly correlated with discrimination based on race. Hence, we conclude that the RR that she found is robust to unobserved confounding, demonstrating a strong association between race and juror selection.[5]

In fact, the inventors of the E-value, Tyler VanderWeele (Professor at Harvard) and Peng Ding (Professor at UC Berkeley) state clearly in their publication (2017):

> "The E-value should be interpreted in the context of the effect sizes that an unmeasured confounder is likely to have with respect to the outcome and treatment. In the context of **biomedical and social sciences research**, effect sizes > 2 or 3 -fold occasionally

[5]Prosecutors have provided the actual reasons they struck most of the Black jurors in the MSU study and these statements provide evidence that disputes Dr. Li's hypothesis that unobserved confounders can explain the prosecutors' strike decisions.

> occur but are not particularly common; a variable that affects both treatment and outcome each by 2- or 3-fold would likely be even less common." (emphasis ours)

Surprisingly, Dr. Li states "in social and behavioral sciences, commonly there is much less information about the causal mechanism between many factors and the observed variables are easily subject to various errors in measuring and collecting. In these cases, there are much more likely some factors that affect the treatment and the outcome differentially by 2 to 3 fold but are not observed." Li, pp 21-22. We have found no support for this assumption and Dr. Li does not back up this assertion with any evidence from the literature.

On the other hand, we found an example in the social sciences where the E-value was calculated to investigate the potential effect of unobserved confounding (Morgan, 2021). This study assessed the impact of racial and ethnic disparities on disability identification among U.S. high school students. In their paper, Morgan calculates E-values to estimate the strength necessary for an unobserved confounder to fully explain the recently reported disability under-identification of students who are Black or Hispanic. They found that the strength necessary for an unobserved confounder to result in their conclusions being reversed would have to be "larger than the risk observed for other factors including biological sex or economic disadvantage (e.g., adjusted ORs 1.64–1.99, or approximately RRs of 1.5–1.73)." In other words, an E-value > 2 would be very unlikely in this context. Morgan concludes that "unmeasured confounding is an unlikely explanation for the observed associations".

4 Conclusions

Here we enumerate our final conclusions.

1. Our results, after careful accounting for important measured confounders and missing data, demonstrate a strong and statistically significant association between race and peremptory strike. In fact, we find that, after controlling for 24 variables, Black venire members are 75% more likely than non-Black members to be struck from juries (RR = 1.75, 95% CI (1.62, 1.89)) (or 58% more likely when death penalty reservations are treated as a race-neutral confounder (RR = 1.58 (1.46, 1.70)).

2. We find that these results are even more notable in cases in which the defendant is Black. In these cases, Black venire members are 94% more

likely than non-Black members to be struck (RR = 1.94, 95% CI (1.75, 2.16)).

3. It is highly unlikely that there are "numerous unmeasured confounders" that would mitigate these findings as Dr. Li asserts. In fact, she only mentions three such factors: appearance, manner, and body language of the prospective juror. We assert that these are not, by definition, confounders of race and strike status and are arguably closely related to racial discrimination.

4. We disagree with Dr. Li's conclusion that the E-value of 2.3 (or our E-value of 2.9) is "not that high". We have not found any evidence to support that, after adjusting for 25 factors, there could exist such a strong hypothetical unobserved factor that would explain away our findings.