

Introduction: AI Bias in Healthcare

- Commercial algorithms that use healthcare expenditures as a proxy for illness severity systematically underestimate need in Black patients, affecting nearly 3 million individuals¹
- A 2019 study by Obermeyer et al. found that resolving this bias would increase the percentage of Black patients receiving crucial care from 17.7% to 46.5%²
- Closed-box models like this are becoming standard in clinical settings, even as they inherit bias from the systemic discrimination embedded in training data^{1,2}

Objective: audit a nationally representative healthcare expenditure database for structural bias and determine whether it can be corrected before bias propagates into downstream models.

Methodology: Bias Auditing Matrix

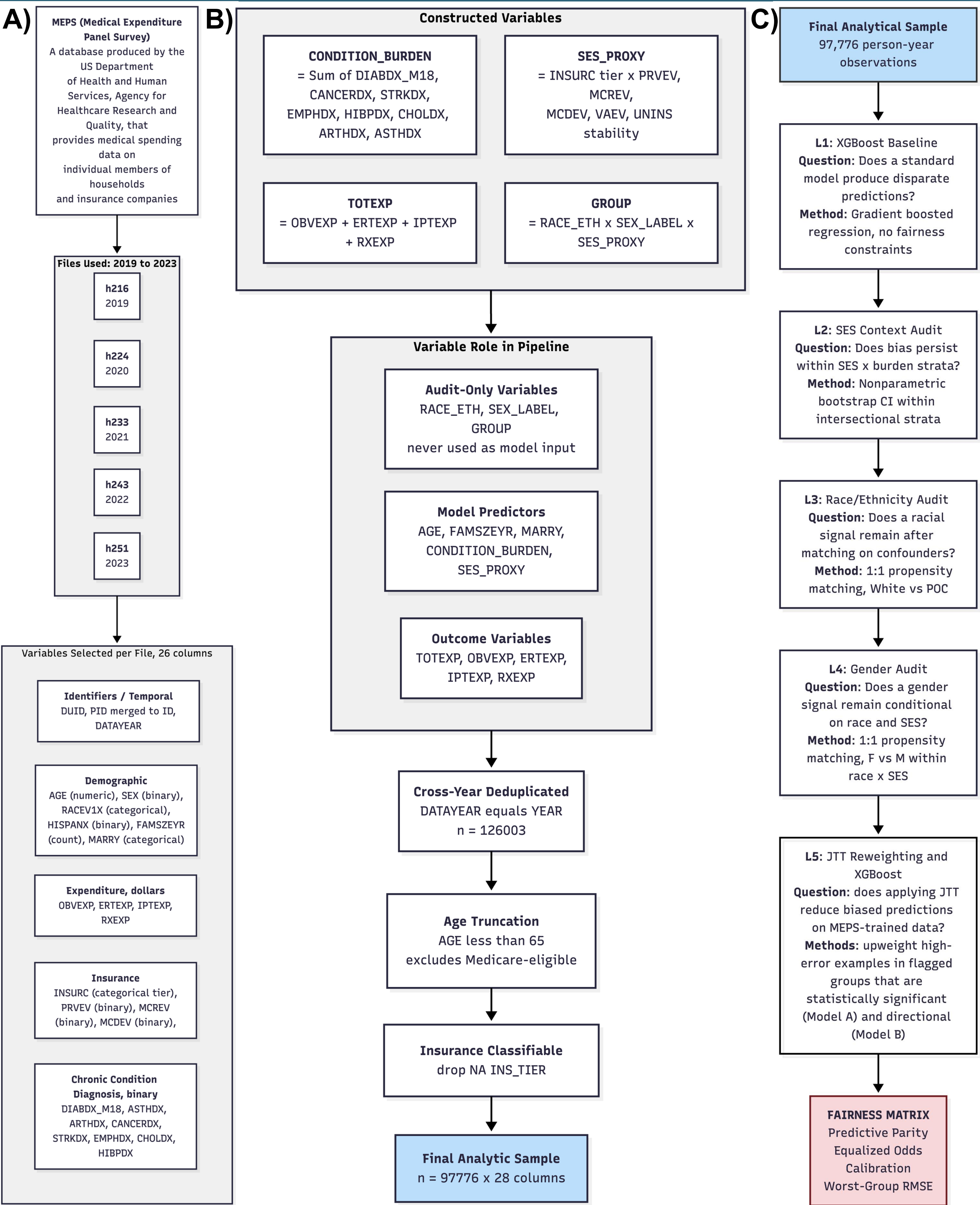
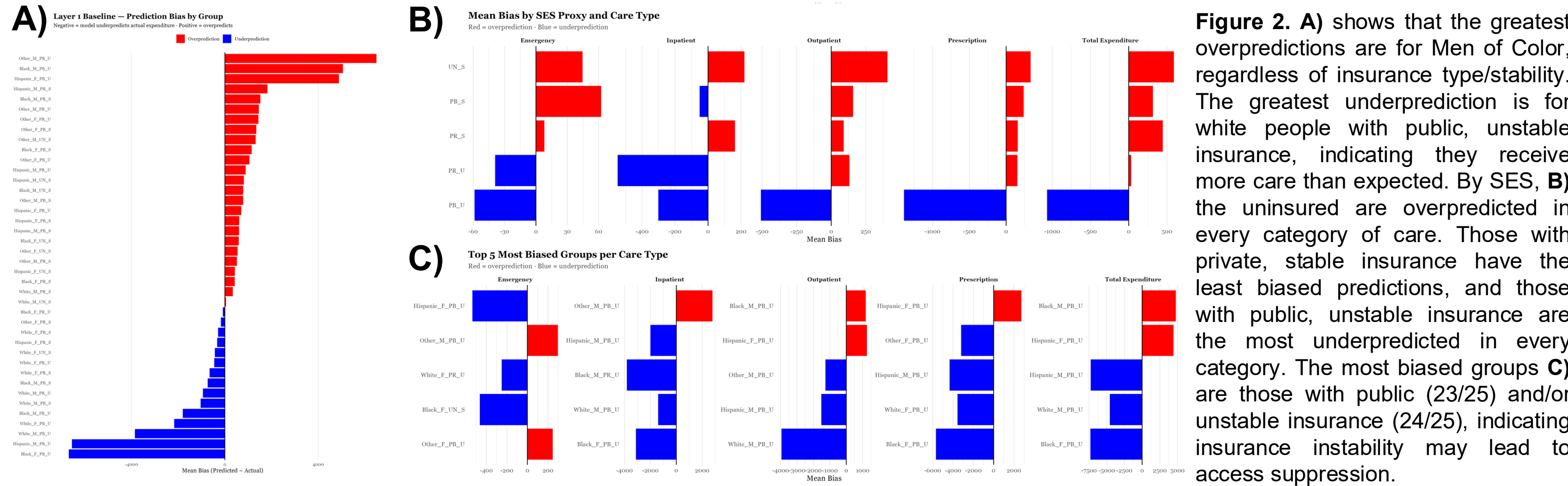


Figure 1. A) describes the MEPS database, data files chosen (2019-2023), and the variables selected by data type and analytical intent.³ **B)** including how they were used to create constructed and audit-only variables, which were later operationalized into predictors and outcome variables that the **C)** L1-L5 pipeline produced a fairness matrix for.

Results: Emergent Disparities and Reweighted Improvements

Baseline Predictions Diverge by Race, SES, and Gender



Disparities Persist for the Uninsured Across Race, Sex, and Condition Burden

- The uninsured have the most widespread disparities (59 v. 42, 39, 22, 13, respectively)
- Indicates persistent differences in outcome by demographic subgroups, meaning the bias is not evenly distributed by race and sex, necessitating the audits below

Figure 3. flags represent a statistically significant bootstrapped bias (95% CI excludes zero). Uninsured patients (UN_S) accumulate 59 flags across all five care types, more than any other SES stratum, indicating the model's error for this group recurs across race, sex, and condition burden. This suggests that SES may play a substantial role in the disparities the model produces. Those with public, unstable insurance (PB_U) show some of the largest point-estimate disparities in **Figure 2**, because PB_U subgroups are frequently small in sample size, producing wide confidence intervals that fail to reach significance despite large observed gaps.

Significant Flags by SES Stratum

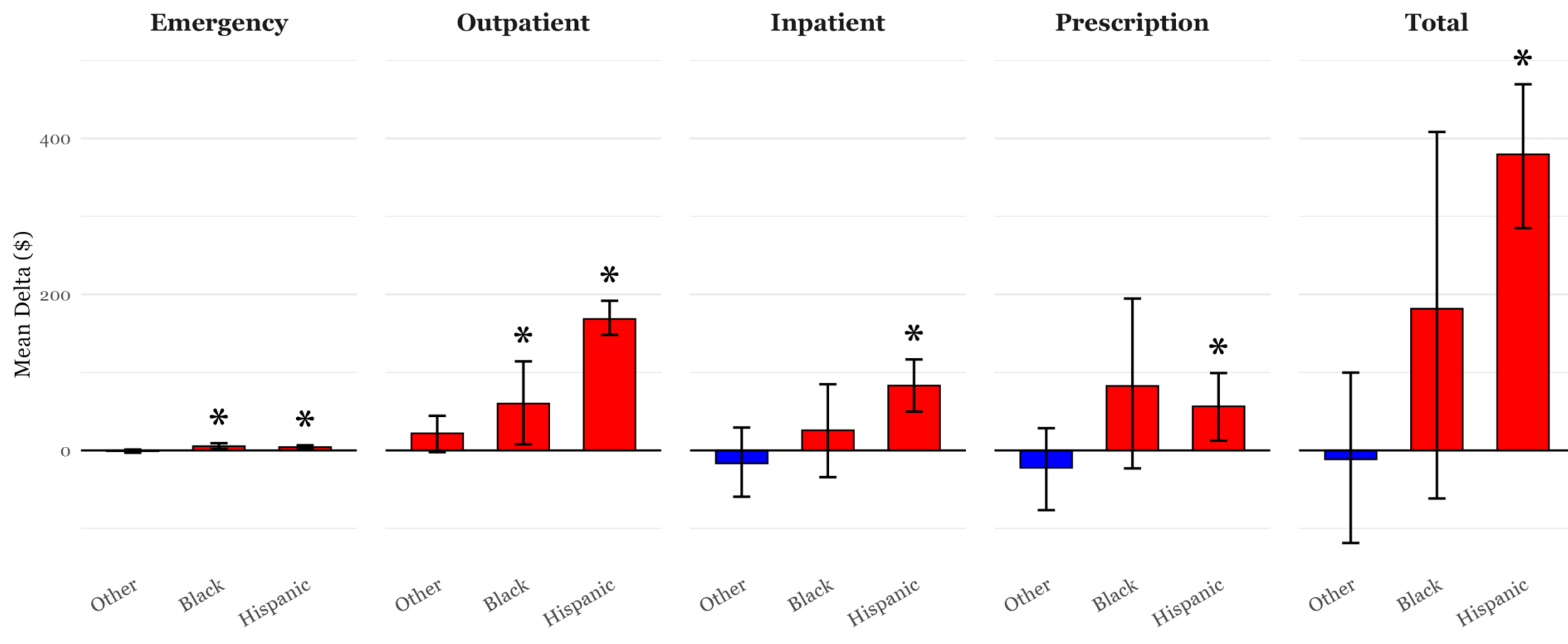
Count of race x sex x burden cells with CI not crossing zero

	ERTEXP	IPTEXP	OBVEXP	RXEXP	TOTEXP
UN_S	8	13	13	15	10
PB_S	9	9	8	9	7
PR_S	3	3	13	10	10
PR_U	4	4	4	8	2
PB_U	1	2	3	4	3

Racial Disparities Persist for Hispanic Patients After Counterfactual Swap

Counterfactual Race Swap: Predicted Expenditure Delta

Red = predicts more for White - Blue = predicts more for POC - Error bars = 95% CI



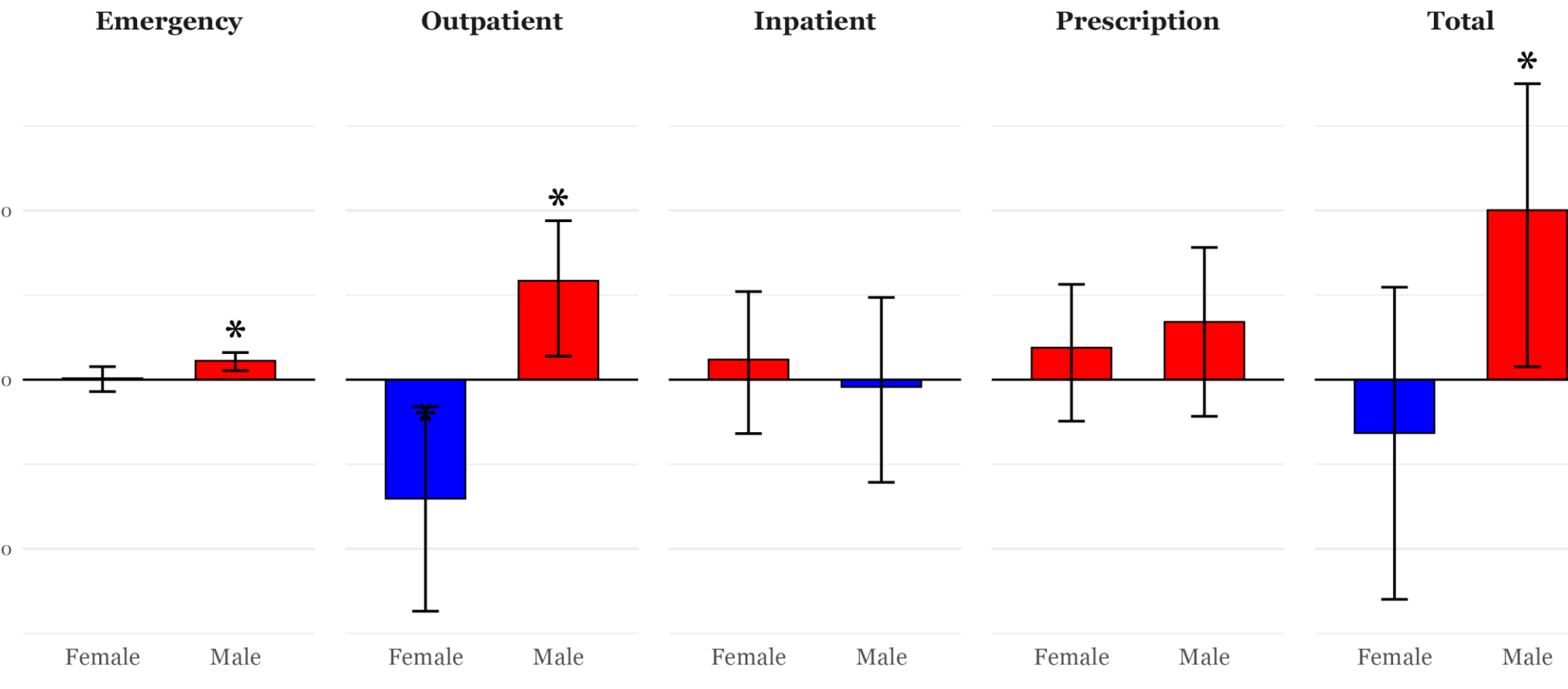
- Black and Hispanic patients are predicted to need less care than white counterparts with the same insurance and condition burden, indicating that racial artifacts persist in the data despite never been encoded as a predictor

Gender Disparity Persists After Matching for Condition Burden and Household

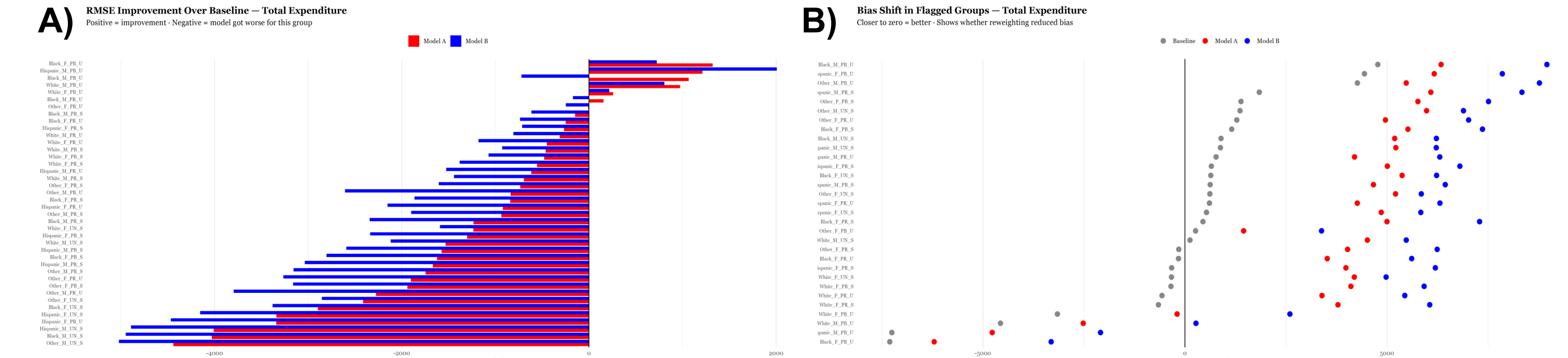
- Men are significantly overpredicted for emergency, outpatient, and total care, indicating less access to care than women with the same household circumstances
- Women are significantly underpredicted for outpatient care, indicating more engagement with preventative care

Bias Within Matched Gender Pairs

Red = overprediction - Blue = underprediction - Error bars = 95% CI - * = significant



Reweighting Cuts Worst-Group Error But Sacrifices Broader Accuracy



Normative Analysis: Rawlsian Justice

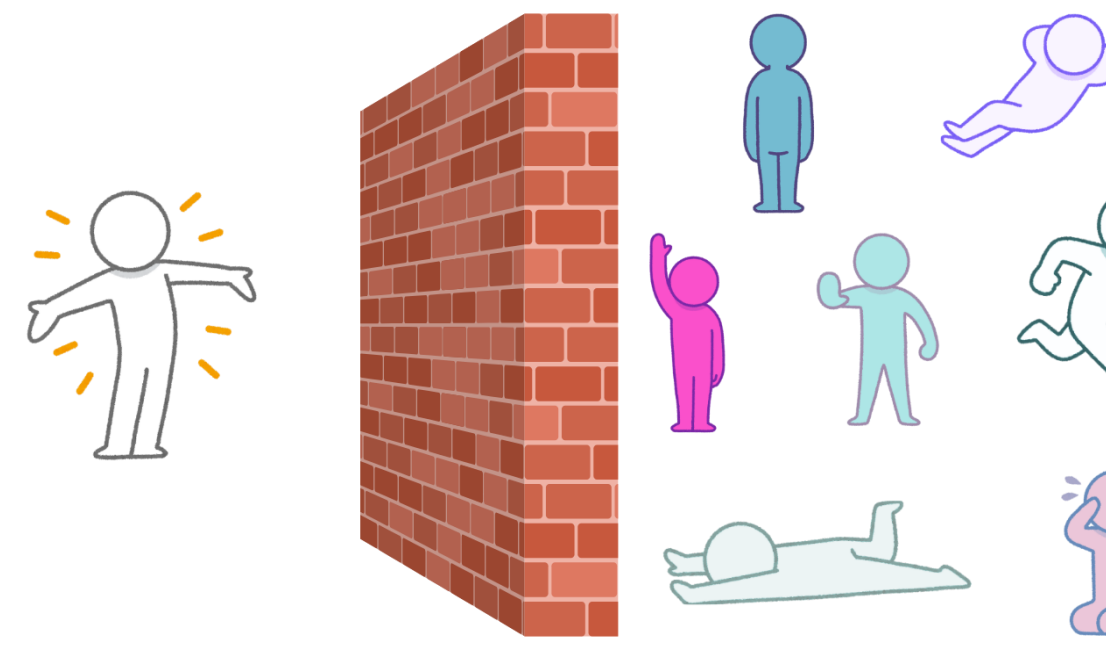


Figure 7. Illustration of Rawls' Veil of Ignorance, Image inspired by ethicalpsychology.com.⁴

- Assuming individuals are born from an original position, behind the Veil of Ignorance, they have a vested interest in the broad wellbeing of all members of a society
- The baseline model systematically produced its highest prediction error for uninsured and coverage-unstable populations, creating a clear subclass by Rawlsian measures
- JTT reweighting improved outcomes for the worst-off groups identified above, while harming broader model performance

- The JTT trade-off is justifiable by the Difference Principle, as Rawls holds that inequality is acceptable when and only when it benefits the worst-off
- If there exists a system with vulnerable populations, it is morally imperative to redesign it to aid the vulnerable

Policy Applications: ACA & Stable Insurance

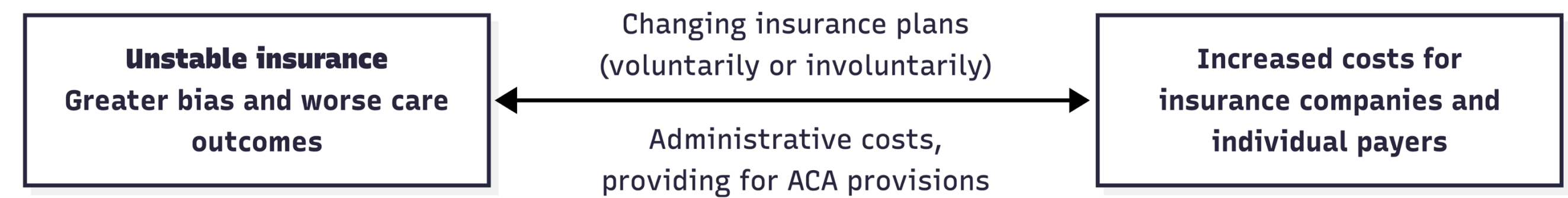


Figure 8. as individuals lose access to healthcare, the financial burden increases for both insurance companies and individual payers. The care burden for individuals without stable insurance also increase, as seen in **Figures 2-6**.

To help address the biases faced by those with unstable insurance, a positive liberty to stable insurance access—for a minimum of 12 consecutive months— is proposed.

- A national public health system has been denied to the American people based on the tradition test, as established in *Washington v. Glucksberg* (1997), despite existing in peer constitutional democracies (Canada, Germany, United Kingdom)⁵
- National Federation of Independent Businesses v. Sebelius* (2012) established that it was within Congressional taxing power to support the ACA through the individual insurance mandate, and provide Congress with the power to fiscally encourage access to healthcare and/or health insurance⁶
- This ruling, with the ACA provisions as outlined in **Figure 8**, provide the legal infrastructure for enforcing a 12-month insurance mandate⁷

Conclusions

- Those with unstable access to health insurance experience biases in healthcare expenditure models that can be improved using JTT, however, the model accuracy for other groups is degraded with these improvements
- Racial and gender artifacts are present in the data, despite never being encoded as predictors, indicating that bias is present in the healthcare system
- To ensure models used commercially retain less bias, databases like MEPS need to be audited and improved

Limitations and Future Directions

- A more in-depth investigation into the experiences of those with public, unstable access should be conducted, as changing access to public insurance (Medicaid), may indicate financial crisis or distinct periods of hardship
- The analysis shown is for the model's internal behavior. To produce findings that are population-representative, person-level weights (PERWT) need to be used
- JTT was trained on total expenditures (TOTEXP) only, future work should investigate by care type
- JTT only shows improvement for the worst-off groups, with some poorly-performing groups experiencing worse outcomes after retraining. Future work should investigate different models that have a more net positive effect

References & Acknowledgements

Thanks to Dr. Bonnie Hurwitz, whose ongoing advocacy and support made this presentation possible; to Dr. Jack Harris pushed for encouraging more pragmatic, critical thinking on philosophy and public policy; and to the Goodnight Scholarships staff, friends, and family also provided valued support.

