Assessing Racial Bias in Clinical Prediction for Urinary Tract Infections

Joshua W. Anderson, MS[1], Nader Shaikh, MD, MPH[2], Shyam Visweswaran, MD PhD[1]
[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA
[2]Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA

Introduction

Race-based clinical prediction tools have the potential to introduce disparities in clinical care. Efforts are underway to carefully investigate the role of race in existing clinical prediction tools and reformulate prediction tools without race. However, such reformulations must be carefully evaluated in terms of both performance and racial bias. We investigated the impact of reformulating UTICalc, a prediction tool designed to help limit catheterizations to children at high risk of urinary tract infection. In response to calls to address racial bias, the original logistic regression (LR) model underlying UTICalc was recently respecified by removing race as a predictor.

UTICalc Version 1 (v1) was released in 2018 and included five features: age, race, gender, maximum temperature, and alternate fever source[1]. In 2022, a respecified UTICalc Version 3 (v3) was released, with the race removed and UTI history and 48-hour fever added as new features[2]. We compared UTICalc v1 and v3 on fairness and discriminative performance metrics to understand how v3 improved over v1.

Methods

We obtained the original datasets and the LR models for UTICalc v1 and v3. UTICalc v1 was trained on a dataset containing 407 blacks and 1,186 nonblack (mostly white) children, and v3 was trained on the same dataset with minor differences (398 blacks and 1,154 nonblacks). Group fairness metrics assess the fairness of two or more groups defined by a sensitive attribute. Using race as the sensitive attribute, we computed two group fairness metrics: demographic parity and equality of odds. Demographic parity requires an equal proportion of positive predictions in blacks and nonblacks, and equality of odds requires that the true positive rate and the false positive rate be equal across the racial groups. To assess the improvement in fairness of UTICalc v3 compared to v1, we compared the difference and ratios of demographic parity and equalized odds[3]. Statistical significance was tested using Wilcoxon rank-sum tests of bootstrapped distributions of these metrics. The optimal value of difference for each metric is 0, and the optimal value of the ratio is 1. To assess changes in discriminative performance, we computed the area under the ROC curve (AUROC), sensitivity, and specificity.

Table 1. Comparison of group fairness metrics for black and nonblack children.

	v1	v3	p-value
Dem. parity diff	0.23	0.05	<0.01
Dem. parity ratio	0.73	0.93	<0.01
Equality of odds diff	0.23	0.03	<0.01
Equality of odds ratio	0.72	0.97	<0.01

Results

Table 1 shows that UTICalc v3 had statistically significant improvement on the fairness metrics. Demographic parity difference and ratios improved from 0.23 to 0.05 and 0.73 to 0.93, respectively. Equalized odds difference and ratios improved from 0.23 to 0.03 and 0.72 to 0.97, respectively. Table 2 shows the AUROC, sensitivity, and specificity values. In UTICalc v3, sensitivity for black children increased by 3.9 percentage points while it only decreased by 2.3 for nonblack children.

Table 2. Comparison of performance metrics for black and nonblack children.

	AUROC		Sensitivity		Specificity	
	v1	v3	v1	v3	v1	v3
All	0.80	0.80	0.97	0.96	0.25	0.35
Black	0.72	0.73	0.90	0.94	0.41	0.34
Nonblack	0.79	0.81	0.98	0.96	0.18	0.36

Discussion and Conclusion

The original study stipulated that the clinical use of UTICalc required a sensitivity of at least 95%. Our results show that the respecified UTICalc model significantly decreased racial bias as measured by the two fairness metrics, and significantly increased sensitivity in blacks while keeping the sensitivity stable in nonblacks. Further evaluation of UTICalc v3 on a different dataset is needed to demonstrate generalizability.

References

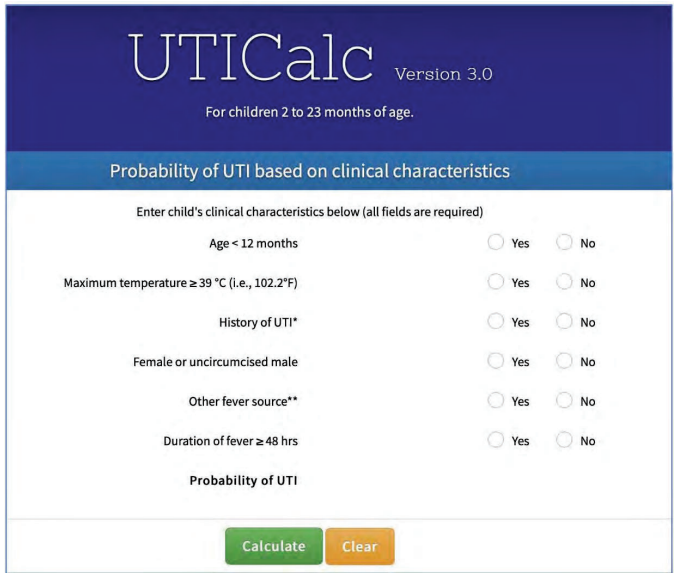
1. Shaikh N, Hoberman A, Hum SW, et. al. Development and validation of a calculator for estimating the probability of urinary tract infection in young febrile children. JAMA Pediatr. 2018 Jun 1;172(6):550-556.
2. Shaikh N, Lee MC, Stokes LR, et. al. Reassessment of the role of race in calculating the risk for urinary tract infection: a systematic review and meta-analysis. JAMA Pediatr. 2022 Jun 1;176(6):569-575.
3. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H, Walker K. Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32. 2020 May.

Assessing Racial Bias in Clinical Prediction for Urinary Tract Infections

Joshua W. Anderson, MS[1], Nader Shaikh, MD, MPH[2], Shyam Visweswaran, MD PhD[1]
[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA
[2]Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA

Background

Race-based clinical prediction tools have the potential to introduce disparities in clinical care. Efforts are underway to carefully investigate the role of race in existing clinical prediction tools and reformulate prediction tools without race. However, such reformulations must be carefully evaluated in terms of both performance and racial bias. We investigated the impact of reformulating UTICalc, a prediction tool designed to help limit catheterizations to children at high risk of urinary tract infection. In response to calls to address racial bias, the original logistic regression (LR) model underlying UTICalc was recently respecified by removing race as a predictor.



Methods

We obtained the original datasets and the LR models for UTICalc v1 and v3. UTICalc v1 was trained on a dataset containing 407 blacks and 1,186 nonblack children (mostly white), and v3 was trained on the same dataset with minor differences (398 blacks and 1,154 nonblacks).

To assess improvement in fairness of UTICalc v3 compared to v1, we compared the difference and ratios of demographic parity and equalized odds. Statistical significance was tested using Wilcoxon rank-sum tests of bootstrapped distributions of these metrics. The optimal value of difference for each metric is 0 and the optimal ratio is 1. To assess changes in discriminative performance, we computed the area under the ROC curve (AUROC), sensitivity and specificity.

Results

Table 1. Comparison of group fairness metrics for black and nonblack children.

	v1	v3	p-value
Dem. parity diff	0.30	0.05	<0.01
Dem. parity ratio	0.73	0.93	<0.01
Equal Opp. diff	0.16	0.03	<0.01
Equal Opp. ratio	0.91	0.97	<0.01
Equality of odds diff	0.28	0.03	<0.01
Equality of odds ratio	0.72	0.97	<0.01

Table 1 shows that UTICalc v3 had statistically significant improvement on fairness metrics. Demographic parity difference and ratios improved from 0.23 to 0.05 and 0.73 to 0.93 respectively. Equal Opportunity improved from 0.16 to 0.03 and 0.91 to 0.97 respectively. Equalized odds difference and ratios improved from 0.23 to 0.03 and 0.72 to 0.97 respectively. Table 2 shows the AUROC, accuracy, sensitivity and specificity values. In UTICalc v3, sensitivity for black children increased by 12 percentage points while it only decreased by 2 for nonblack children.

Table 2. Comparison of performance metrics for black and nonblack children.

	AUROC	Accuracy	Sensitivity	Specificity
UTICalc v1				
Overall (N=1575)	0.64	0.55	0.94	0.35
Black (N=407)	0.68	0.59	0.82	0.59
Nonblack (N=1186)	0.62	0.54	0.98	0.26
UTICalc v3				
Overall (N=1562)	0.65	0.56	0.95	0.35
Black (N=408)	0.64	0.46	0.94	0.34
Nonblack (N=1154)	0.66	0.59	0.96	0.36

Discussion and Conclusion

The respecification of the UTICalc model significantly removed racial bias. The improvement is mostly explained by a large increase in sensitivity for black children while keeping the nonblack sensitivity stable. The original study emphasizes the importance of sensitivity, stating that most clinicians would require a minimum of 95%. UTICalc v3 significantly reduced racial bias, though it could be improved further with additional data and bias mitigation methods.

References

1. Shaikh N, Hoberman A, Hum SW, et. al. Development and validation of a calculator for estimating the probability of urinary tract infection in young febrile children. JAMA Pediatr. 2018 Jun 1;172(6):550-556.
2. Shaikh N, Lee MC, Stokes LR, et. al. Reassessment of the role of race in calculating the risk for urinary tract infection: a systematic review and meta-analysis. JAMA Pediatr. 2022 Jun 1;176(6):569-575.
3. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H, Walker K. Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32. 2020 May