Evaluating an Artificial Intelligence (AI) Communication Platform for Racial Biases in Information Quality about Prostate Cancer (PCa) Germline Testing

Johanna Balas, MD[1], Lauren Bowling, MS, CGC[2], Nonna Shakhnazaryan, BS[3], Rahul R. Aggarwal[3,4], MD, Hala T. Borno, MD[3], Julian C. Hong, MD, MS[3,5], Franklin W. Huang, MD, PhD[1,3], Barry Tong, MS, CGC[2], Daniel H. Kwon, MD[3]

Affiliations:
1. School of Medicine, University of California, San Francisco, San Francisco, CA, USA
2. Cancer Genetics and Prevention Program, University of California, San Francisco, San Francisco, CA, USA
3. Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA
4. Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, CA, USA
5. Department of Radiation Oncology, University of California, San Francisco, San Francisco, CA, USA

Background: Guidelines recommend germline testing in advanced PCa to inform treatment and personal/familial cancer risk, but Black patients are less likely to complete testing than white men. AI communication platforms are increasingly being incorporated in pre-test counseling to augment patient education, access, and equity. However, these platforms have not been evaluated for racial bias, which is a known hazard in other clinical contexts.

Methods: We specialized a private, secure generative AI platform developed by UCSF for pre-test PCa genetics education, called "ProGene". We asked ProGene 7 frequently asked questions (FAQs): types of genetic testing and test results, personal benefits of testing, familial benefits, drawbacks, logistics, costs, and privacy concerns. We asked each question with a standardized prompt 9 times: three simulated patients with metastatic PCa (Black, non-Hispanic White, and race-agnostic), each in triplicate, for a total of 63 questions. Two blinded reviewers assessed the 63 responses across 4 information quality domains: 1) Comprehensiveness (mean, 0–100%) using an investigator-created rubric, 2) Accuracy (proportion, presence/absence of any inaccurate statement), 3) Readability (mean grade level) via SMOG and Flesch-Kincaid formulas, and 4) Actionability (mean, 0-100%) based on the Patient Education Materials Assessment Tool. We used the two-sample t-test and Wilcoxon rank-sum test to compare continuous outcomes, and chi-squared test for categorical outcomes, between race subgroups overall and for each FAQ.

Results: Table 1 summarizes outcomes by race. For comprehensiveness, the overall mean score was 67% and did not vary by race. However, for FAQ 1 (types of genetic testing and test results), ProGene responses to the Black patient were less comprehensive than those to the race-agnostic patient (60% vs 93%; p=0.008). Inaccuracies were present in 32% of responses and did not vary by race. Mean readability was 10th (SMOG) and 13th grades (Flesch-Kincaid) and did not vary by race. Actionability was 92% and did not vary by race.

Conclusion: We did not identify major racial biases in the quality of the AI communication platform's responses to PCa germline testing questions. Overall actionability was high, comprehensiveness and accuracy were moderate, but readability was limited, presenting opportunities for improved prompt engineering, AI model updates, and human oversight. AI communication platforms are a promising tool to promote equity in PCa germline testing delivery, warranting continued refinement and evaluation.

Funding: Prostate Cancer Foundation #24YOUN09.

Conflicts of interest: The authors have no relevant conflicts of interest to disclose.

Table 1. Evaluation of AI Platform Response Quality by Race

Quality domain	Black	Non-Hispanic white	Race agnostic
Comprehensiveness (%)	64%	63%	74%
Inaccuracy rate (%)	24%	38%	33%
Readability (grade level)	11	12	12
Actionability (%)	90%	92%	93%