## ACS 2025 Surgeons and Engineers: A Dialogue on Surgical Simulation Meeting

P-E-08

## **Research Abstracts**

## Quantifying the Effect of Implicit Bias on Performance of Unbiased Deep Learning Models for Surgical Skill Assessment

Nanthini Narayanan, MSE; Divyasree Sasi Kumar, MSE; Vishal Patel, PhD; Shameema Sikder, MD, FACS; and S. Swaroop Vedula, MBBS, PhD, MPH

Johns Hopkins University, Baltimore, MD; Wilmer Eye Institute, Johns Hopkins University, Baltimore, MD; Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD

**Introduction:** Video-based surgical skill assessment (VBA) is crucial for surgical education and improved patient outcomes. Implicit bias in raters is known to affect evaluation of surgeons but the impact of implicit bias on artificial intelligence (AI) models for VBA is not well studied. Specifically, while it is intuitive that training AI models with biased ground truth will lead to biased predictions, the effect of bias in test data on performance of AI models is less evident. Our objective in this study was to quantify how much implicit bias in the test data affects performance of unbiased AI models for VBA.

**Methods:** We used 99 videos of capsulorhexis from one institution, Dataset-99, and 97 videos from the publicly available Cataract-101 dataset. An expert surgeon provided unbiased skill ratings using a standard scale. We trained AI models for VBA with the unbiased ground truth in each dataset. We tested the models with biased datasets created by injecting implicit bias, in increments of 10%, favoring female and male surgeons. We computed the area under the receiver operating characteristic curve (AUC) of the unbiased models on biased test datasets.

**Results:** The AUC of AI models on unbiased test data was 0.8 (95% confidence interval [CI], [0.71,0.89]) in Dataset-99 and 0.62 (95% CI, [0.51,0.74]) in Cataract-101. With biased test data, for every 10% increase in intensity of implicit bias, the mean AUC reduction was 2.5% when the bias favored women, 9.49% when the bias favored men, and 6% overall in Dataset-99. The corresponding estimates for Cataract-101 were 1.89%, 0.81%, and 1.35%, respectively.

**Conclusions:** Even with unbiased AI models for VBA, their performance decreased as the intensity of bias in test data increased. Datasets used to evaluate AI for VBA must address the adverse impact of implicit bias.

