Evaluating the Racial and Ethnic Equity of Machine Learning Algorithms in Predicting 30-Day Complications Following Total Hip and Knee Arthroplasty

Christian A. Pean MD[1], Anirudh Buddhiraju MD[1], Henry Hojoon Seo BA[1], Michelle Shimizu BSc[1], Tony Lin-Wei Chen MD, PhD[1], MohammadAmin RezazadehSaatlou MD[1], Ziwei Huang MD, PhD[1], Shane Fei Chen MA[1], Blake M. Bacevich BSc[1], Oh-Jak Kwon[1], Jona Kerluku BSc[1], John G. Esposito MD[1], Young-Min Kwon MD, PhD[1]

[1]Bioengineering Laboratory, Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA
ymkwon@mgh.harvard.edu

Disclosures: Christian A. Pean (N), Anirudh Buddhiraju (N), Henry Hojoon Seo (N), Michelle Shimizu (N), Tony Lin-Wei Chen (N), MohammadAmin RezazadehSaatlou (N), Ziwei Huang (N), Shane Fei Chen (N), Blake M. Bacevich (N), Oh-Jak Kwon (N), Jona Kerluku (N), John G. Esposito (N), Young-Min Kwon (5- MicroPort; 5- Depuy; 5- Smith & Nephew; 5- Stryker; 5- Zimmer Biomet)

INTRODUCTION: Machine Learning (ML) algorithms have emerged as promising tools to risk-stratify patient populations and improve care coordination. Although the predictive capabilities of ML algorithms for total hip arthroplasty and total knee arthroplasty have been demonstrated in previous studies, the performance of these tools in racial and ethnic minority patients has not been investigated. With the increasing use of machine learning tools for risk stratification of patients, it is important to ensure that the predictive ability of ML algorithms is equitable. The aim of this study was to assess the performance of ML algorithms in predicting 30-day complications following hip or knee total joint arthroplasty (TJA) in racial and ethnic minority patients.

METHODS: 267,194 patients who underwent primary TJA between 2013 and 2020 in the United States were identified from the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database. The patient cohort was stratified according to race as White, African American or Black, Asian, American-Indian, or Native Hawaiian, with further sub-stratification into Hispanic or non-Hispanic ethnicity. Histogram-based gradient (HGB) boosting and random forest (RF), were modeled to predict 30-day complications following primary TJA in the overall population and were assessed using their ability to distinguish between patients who experienced 30-day complications and those who did not, calibration, and potential clinical usefulness. Both models were evaluated on their predictive performance in each racial and ethnic sub-cohort and compared across all groups.

RESULTS: Both ML models achieved excellent discrimination (AUC>0.8) in the non-Hispanic white population (AUC_{HGB}=AUC_{RF}=0.86), with excellent calibration (HGB: slope= 1.00, intercept= -0.03, Brier score= 0.12; RF: slope= 0.97, intercept= 0.02, Brier score= 0.12). Model discrimination decreased in the White Hispanic (AUC: 0.75–0.76), Black (AUC=0.77), Black Hispanic (AUC=0.78), Asian non-Hispanic (AUC: 0.78–0.79), and overall (AUC: 0.75–0.76) cohorts (**Fig. 1**) but remained well-calibrated (**Fig. 2**). Model discrimination (AUC: 0.67–0.68) and calibration were the least in the American Indian cohort (**Table 1**).

DISCUSSION: This study demonstrated that machine learning algorithms were not equitable in their predictive ability for 30-day complications following primary THA and TKA in racial and ethnic minorities. Our findings underscore the importance of developing inclusive machine learning models for preoperative risk stratification of patients to protect healthcare equity. As the use of ML algorithms expands, support for an equity-conscious context for underserved patient populations is necessary.

SIGNIFICANCE/CLINICAL RELEVANCE: Machine learning algorithms are not equally effective in predicting 30-day complications following primary total hip and total knee arthroplasty in racial and ethnic minorities.

Table 1. Discrimination and calibration of HGB and RF machine learning algorithms for the prediction of 30-day complications following primary TJA in different racial and ethnic subpopulations

Algorithm	AUC	Calibration slope	Calibration intercept	Brier score
Overall				
HGB	0.75	0.72	0.05	0.16
RF	0.76	0.86	0.03	0.15
White				
HGB	0.86	1.00	-0.03	0.12
RF	0.86	0.97	0.02	0.12
White Hispanic				
HGB	0.75	0.70	0.11	0.18
RF	0.76	0.82	0.12	0.18
African American				
HGB	0.77	0.93	0.01	0.17
RF	0.77	0.89	0.05	0.17
African American Hispanic				
HGB	0.78	1.08	0.01	0.17
RF	0.78	1.12	0.03	0.17
Asian				
HGB	0.79	0.88	0.05	0.17
RF	0.78	0.73	0.13	0.17
American Indian				
HGB	0.68	0.40	0.009	0.19
RF	0.67	0.35	0.05	0.18
Native Hawaiian				
HGB	0.84	0.63	0.05	0.13
RF	0.84	0.81	0.005	0.12

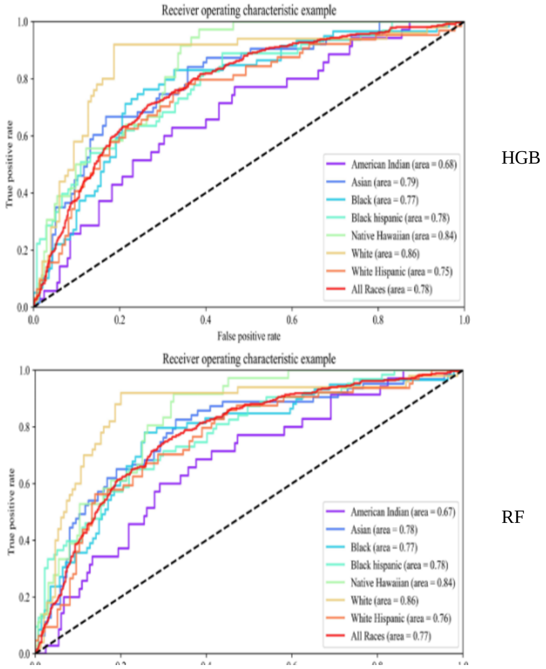

Fig 1. Receiver operating characteristic curves of the histogram-based gradient boosting and random forest algorithms for the prediction of 30-day complications following primary total hip or total knee arthroplasty in different racial and ethnic subpopulations.