Machine Learning Can Mitigate Racial Bias in Orthopaedics: Predicting Overnight Stay After Knee Arthroscopy

Jonathan S. Lee[1], Bilal S. Siddiq[1], Stephen M. Gillinov[1], Kieran S. Dowley[1], Megan J. Wei[2], Nathan J. Cherian[1], Mark P. Cote[1], Scott D. Martin[1]

[1]Massachusetts General Hospital Division of Sports Medicine, Boston, MA, [2]Johns Hopkins Department of Computer Science, Baltimore, MD

Email of Presenting Author: jlee376@mgh.harvard.edu

INTRODUCTION: Over the past 3 years, the utilization of machine learning in orthopaedics has risen exponentially. Past studies have employed machine learning to predict hospital readmission, rates of achieving clinical thresholds, and functional improvements. Due to the utility of these models, some orthopaedic researchers have gone as far as publishing these algorithms through online calculators, allowing users to input pertinent patient data and derive a predicted clinical outcome. Despite machine learning's immense potential for clinical application, before researchers can offer these tools to a broad and diverse patient population, they must first ensure that their machine learning model performs with high degrees of accuracy irrespective of patient race. The purpose of the present study was to develop a racially equitable machine learning model that accurately predicts overnight stay following arthroscopic knee surgery.

METHODS: This retrospective study queried the NSQIP database for patients ≥ 18 years old who underwent knee arthroscopy between 2011 and 2020. Patients who self-identified as White, Black or African American, Asian, Native American or Alaskan Native, or Native Hawaiian or Pacific Islander and had complete demographic, intraoperative, and post-operative outcome data were included in our analysis. All model development and statistical analysis was carried out in Python using packages from Scikit-learn. Predictive analysis was performed using a Random Forest (RF) and performance was optimized through feature selection, hyperparameter tuning with cross validation, and over- and under-sampling techniques with SMOTE. The following patient variables to train the RF model included but were not limited to sex, age, BMI, wound classification, ASA class, and type of anesthesia. Patients with a hospital length of stay (LOS) ≥ 1 day were placed into the overnight stay cohort, while patients with a LOS < 1 represented the same-day discharge cohort. Model performance was assessed using AUC, accuracy, and F1 score.

RESULTS: Of the 75,775 patients who met inclusion criteria, 7.5% (n=5,714) required overnight stay. When stratified by racial self-identification, most patients identified as White (78.3%), followed by Black or African American (13.7%), Asian (5.4%), Native American or Native Alaskan (1.5%), and Native Hawaiian or Pacific Islander (1.1%) (**Table 1**). Despite a lack of representation among minority patient groups, the racial-equitably trained RF model showed high levels of performance for AUC (mean = 0.94 ± 0.03) and accuracy (mean = 0.97 ± 0.01) irrespective of patient race. F1 score, however, varied with the model performing best for Asian (0.92), Native Hawaiian or Pacific Islander (0.86), and Black or African American patients (0.78) (**Figure 1)**. Of the variables included to train the RF model, 5 features accounted for 73% of model development (**Table 2**).

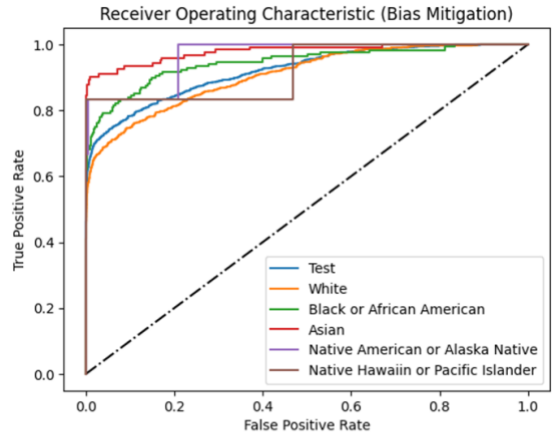
DISCUSSION: We developed a racially equitable machine learning model that effectively predicted overnight stay in patients undergoing knee arthroscopy. As the utilization of machine learning in orthopaedics continues to rise, it is crucial that models intended for clinical application are generalizable and exhibit high levels of performance for minority patients.

SIGNIFICANCE/CLINICAL RELEVANCE: The onus is on orthopaedic researchers to assess the internal validity of their machine learning models before offering them to a broader patient population. Failing to do so may result in inequitable clinical models that exacerbate poor health outcomes among minority patients.

Table 1. Patient Demographics

	Patients (n=75,775)
Same-Day Discharge	5,714 (7.5%)
Overnight Stay	70,061 (92.5%)
Age, years	37.2 ± 14.4
BMI, kg/m^2	29.5 ± 7.3
Sex	
Male	32,101 (42.4%)
Female	43,674 (57.6%)
Racial Self-Identification	
White	59,313 (78.3%)
Black or African American	10,359 (13.7%)
Asian	4,099 (5.4%)
Native American or Alaska Native	1,152 (1.5%)
Native Hawaiin or Pacific Islander	852 (1.1%)
Current Procedural Terminology	
29875	5,291 (7.0%)
29876	6,819 (9.0%)
29877	15,361 (20.3%)
29882	9,025 (11.9%)
29888	39,279 (51.8%)

29875 (arthroscopy knee, surgical; synovectomy, limited)
29876 (arthroscopy, knee, surgical; synovectomy, major, 2 or more compartments)
29877 (arthroscopy knee, surgical; debridement/shaving of articular cartilage)
29882 (arthroscopy, knee surgical; with meniscal repair medial or lateral)
29888 (arthroscopically aided anterior cruciate ligament repair/augmentation or reconstruction)



	AUC	Accuracy	F1 Score
Test	0.92	0.97	0.74
White	0.91	0.96	0.70
Black or African American	0.94	0.97	0.78
Asian	0.98	0.98	0.92
Native American or Alaska Native	0.96	0.99	0.67
Native Hawaiin or Pacific Islander	0.92	0.98	0.86

AUC = Area under the ROC Curve, ROC = Receiver operating characteristic

Table 2. Top variables contributing to model development

Variables	Contribution
Arthroscopic knee procedure	36%
Admission quartile	16%
Anesthesia type	13%
ASA class	5%
Days from hospital admission to operation	3%