“Fairness” Analysis of Machine Learning Models Predicting Readmission and Prolonged Length of Stay Following Total Knee Arthroplasty in Underrepresented Populations

Michelle Riyo Shimizu, Marium Raza, Pengwei Xiao, zhijun li, Anisha Elizabeth Gemmy, William T Sampson, Isaiah Freeman, Sina Afzal, Young-Min Kwon

INTRODUCTION:

While machine learning (ML) models demonstrate high predictive accuracy, recent studies reveal that they underperform for racial and ethnic minorities, suggesting inherent biases that exacerbate health disparities. Assuming a “one-size-fits-all” approach perpetuates inequities in decision-making for marginalized groups. This study aims to assess the “fairness” of validated ML models for total knee arthroplasty (TKA) outcomes and explore bias mitigation strategies to enhance equitable prediction performance using a large national database.

METHODS:

Four ML models were developed and validated using data from 267,428 and 267,413 patients in the ACS-NSQIP database to predict readmission and prolonged length of stay (pLOS; ≥3 days) following TKA, respectively. Bias assessment was performed using protected attributes, which are characteristics that are protected against discrimination (age, gender, race, ethnicity, and diabetes), and various fairness metrics (Table 1). The acceptable fairness metric range was set between 0.8 and 1.25. Three bias mitigation strategies (one postprocessing algorithm and two reduction algorithms) developed by an open-source toolkit were incorporated and trialed for each algorithm. Postprocessing algorithms transform already-trained models by adjusting decision thresholds across different subgroups while optimizing prediction accuracy. Reduction algorithms achieve the same effect by reweighting data points and retraining the developed models. The most effective strategy for a given protected attribute was integrated with the original model and reassessed based on fairness metrics.

RESULTS:

28,490 (10.7%) experienced a pLOS and 8,710 (3.3%) had a 30-day readmission (Table 2). Prolonged length of stay was recorded in 1,655 (13.0%) Hispanic/Latinx patients and 5,682 (16.1%) non-White patients. In contrast, 26,835 (10.5%) non-Hispanic/Latinx and 22,808 (9.8%) White patients experienced a pLOS. 3.4% of Hispanic/Latinx patients (n=439) and 3.4% non-White patients (n=1,214) had a 30-day readmission, comparable to the 3.2% and 3.3% of non-Hispanic/Latinx and White patients, respectively. Random forest (RF) had the best predictive performance ($readmission_{AUC}$ = 0.98; $pLOS_{AUC}$ = 0.92). Inferior predictive equality ratio (PER) for readmission was demonstrated in females (0.47 vs. males), non-White (0.75 vs. White), and Hispanic/Latinx (0.53 vs. not Hispanic/Latinx) subcohorts (Table 3). Significant differences in all but accuracy equality metrics for pLOS were unveiled across all underprivileged cohorts (Table 4). Various bias mitigation strategies were effective in both ML models; however, trade-offs were observed between the fairness metrics (Table 5, 6).

DISCUSSION AND CONCLUSION:

Our study highlights the performance bias of currently validated ML models in predicting readmission and pLOS after TKA across marginalized groups. While non-White and Hispanic/Latinx cohorts had similar or higher rates of readmission and pLOS compared to White and non-Hispanic/Latinx patients, the ML models failed to predict high-risk individuals within these underrepresented cohorts. This disparity is reflected in the lower PER among marginalized groups, which indicates a lower false positive rate—and consequently, a higher false negative rate. A high false negative rate in a model suggests that more at-risk patients go unrecognized, further exacerbating inequity in care. Given high healthcare utilization costs and suboptimal patient outcomes associated with unplanned readmission and pLOS, emphasis should be placed on minimizing the risk of failing to identify at-risk patients. Despite improvement in model performance with bias mitigation strategies, the persistent trade-offs in prediction metrics underscore the need for deliberate model auditing and corrections before these tools are adopted into clinical practice.

Table 1. Overview of fairness metrics.

Fairness Metric	Concept	Performance metric
Equal Opportunity (EOR)	Actual positive will be labeled positive	True Positive
Predictive Equality (PER)	Actual negative will be labeled positive	False Positive
Predictive Parity (PPR)	Positive label is truly positive	Positive Predictive Value
Statistical Parity (SPR)	Probability of positive prediction for both cohorts	Statistical Parity
Accuracy Equality (AER)	Number of true predictions to number of total predictions	Accuracy

Table 2: Demographics by subcohorts of protected attributes and outcomes following primary total knee arthroplasty

	LOS (n=267,428)		Readmission (n=267,413)	
	Normal LOS (n=; %)	pLOS (n=; %)	No Readmission (n=; %)	30-day Readmission (n=; %)
Gender				
Male	146,741 (54.9)	18,796 (7.0)	160,621 (60.1)	4,907 (1.8)
Female	92,197 (34.5)	9,694 (3.6)	98,082 (36.7)	3,803 (1.4)
Age				
≤ 65 years	107,390 (40.2)	11,249 (4.2)	115,450 (43.2)	3,180 (1.2)
> 65 years	131,548 (49.2)	17,241 (6.4)	143,253 (53.6)	5,530 (2.1)
Race				
White	209,260 (78.2)	22,808 (8.5)	224,560 (84.0)	7,496 (2.8)
Non-White	29,678 (11.1)	5,682 (2.1)	34,143 (12.8)	1,214 (0.5)
Ethnicity				
Not Hispanic/Latinx	227,848 (85.2)	26,835 (10.0)	246,399 (92.1)	8,271 (3.1)
Hispanic/Latinx	11,090 (4.1)	1,655 (0.6)	12,304 (4.6)	439 (0.2)
Diabetes				
No Diabetes	41,784 (15.6)	7,045 (2.6)	46,780 (17.5)	2,047 (0.8)
Diabetes	197,154 (73.7)	21,445 (8.0)	211,923 (79.2)	6,663 (2.5)

LOS= length of stay; pLOS=prolonged length of stay

Table 3. Assessment of model fairness and disparity in predicting 30-day readmission following primary total knee arthroplasty.

Attribute	Sub cohort	AER	PER	EOR	PPR	SPR
Gender	Male	1.00	0.76	0.93	1.04	0.89
	Female	1.01	**0.47**	1.02	1.13	0.90
Age	≤ 65 years	1.01	0.55	1.01	1.11	0.91
	> 65 years	1.00	1.06	0.95	1.09	0.88
Race	White	1.01	0.63	1.01	1.09	0.93
	Non-White	1.00	**0.75**	0.93	1.05	0.89
Ethnicity	Not Hispanic/Latinx	1.01	0.50	1.02	1.12	0.91
	Hispanic/Latinx	1.00	**0.53**	0.88	1.09	0.81
Diabetes	No Diabetes	1.01	0.58	1.01	1.10	0.92
	Diabetes	1.00	0.80	0.89	1.02	0.87

Value outside of 0.8 and 1.25 highlighted in bold to demonstrate biased for that metric.
AER= accuracy equality ratio; PER=predictive equality ratio; EOR=equal opportunity ratio; PPR= predictive parity ratio; SPR=statistical parity ratio

Table 4. Assessment of model fairness and disparity in predicting prolonged length of stay following primary total knee arthroplasty.

Attribute	Sub cohort	AER	PER	EOR	PPR	SPR
Gender	Male	0.98	1.03	0.74	0.85	0.84
	Female	0.98	1.15	**0.65**	**0.74**	0.83
Age	≤ 65 years	0.99	0.97	0.72	0.86	0.81
	> 65 years	0.98	1.05	**0.58**	**0.73**	**0.76**
Race	White	1.00	0.66	0.78	1.05	0.72
	Non-White	0.96	**1.27**	**0.38**	**0.45**	**0.72**
Ethnicity	Not Hispanic/Latinx	0.99	0.79	0.66	0.91	0.70
	Hispanic/Latinx	0.97	1.15	**0.55**	**0.67**	**0.78**
Diabetes	No Diabetes	0.99	0.97	0.73	0.87	0.81
	Diabetes	0.97	**0.74**	**0.34**	**0.62**	**0.48**

Value outside of 0.8 and 1.25 highlighted in bold to demonstrate biased for that metric.
AER= accuracy equality ratio; PER=predictive equality ratio; EOR=equal opportunity ratio; PPR= predictive parity ratio; SPR=statistical parity ratio

Table 5. Assessment of model fairness and disparity in predicting 30-day readmission following primary total knee arthroplasty after bias mitigation.

Attribute	Sub cohort	AER	PER	EOR	PPR	SPR
Gender	Male	1.00	0.76	0.93	1.04	0.89
	Female[1]	1.01	0.94	0.80	1.04	1.06
Age	≤ 65 years	1.01	0.55	1.01	1.11	0.91
	> 65 years[1]	1.00	1.20	1.01	1.01	1.18
Race	White	1.01	0.63	1.01	1.09	0.93
	Non-White[1]	0.98	1.06	0.96	0.88	0.89
Ethnicity	Not Hispanic/Latinx	1.01	0.50	1.02	1.12	0.91
	Hispanic/Latinx[2]	0.93	1.24	0.87	**0.67**	0.81
Diabetes	No Diabetes	1.01	0.58	1.01	1.10	0.92
	Diabetes[2]	1.12	0.81	1.25	0.99	0.90

Value outside of 0.8 and 1.25 highlighted in bold to demonstrate biased for that metric.
[1] Utilized ThresholdOptimizer postprocessing algorithm for equalized odds
[2] Utilized RandomForest Classifier reduction algorithm
AER= accuracy equality ratio; PER=predictive equality ratio; EOR=equal opportunity ratio; PPR= predictive parity ratio; SPR=statistical parity ratio

Table 6. Assessment of model fairness and disparity in predicting prolonged length of stay following primary total knee arthroplasty after bias mitigation.

Attribute	Sub cohort	AER	PER	EOR	PPR	SPR
Gender	Male	0.98	1.03	0.74	0.85	0.84
	Female[1]	0.99	1.06	0.99	0.99	1.00
Age	≤ 65 years	0.99	0.97	0.72	0.86	0.81
	> 65 years[1]	1.01	1.01	1.02	1.03	1.09
Race	White	1.00	0.66	0.78	1.05	0.72
	Non-White[1]	0.98	1.06	0.97	0.98	0.99
Ethnicity	Not Hispanic/Latinx	0.99	0.79	0.66	0.91	0.70
	Hispanic/Latinx[1]	0.92	1.08	0.85	**0.76**	**0.66**
	Hispanic/Latinx[2]	1.04	1.16	1.19	**0.68**	1.04
	Hispanic/Latinx[3]	0.93	1.12	0.87	0.81	**0.68**
Diabetes	No Diabetes	0.73	0.97	0.73	0.87	0.81
	Diabetes[1]	1.09	0.89	1.09	0.94	0.83

Value outside of 0.8 and 1.25 highlighted in bold to demonstrate biased for that metric.
[1] Utilized ThresholdOptimizer postprocessing algorithm for equalized odds
[2] Utilized DecisionTree Classifier reduction algorithm
[3] Utilized RandomForest Classifier reduction algorithm
AER= accuracy equality ratio; PER=predictive equality ratio; EOR=equal opportunity ratio; PPR= predictive parity ratio; SPR=statistical parity ratio