Machine Learning for Healthcare 2020 – Clinical Abstract Track

Effects of Mislabeled Race Categorizations on Prediction of Inpatient Hyperglycemia

Morgan Simons[*1,2], *Kristin Corey*[*1,2], *Marshall Nicols*[2], *Michael Gao*[2], , *Suresh Balu*[2], *Mark Sendak*[*2] *Joseph Futoma*[*3,4]

[1] *Duke School of Medicine* [2] *Duke Institute for Health Innovation* [3]*Harvard School of Engineering & Applied Sciences*

[4]*Duke Statistical Science*

**contributed equally*

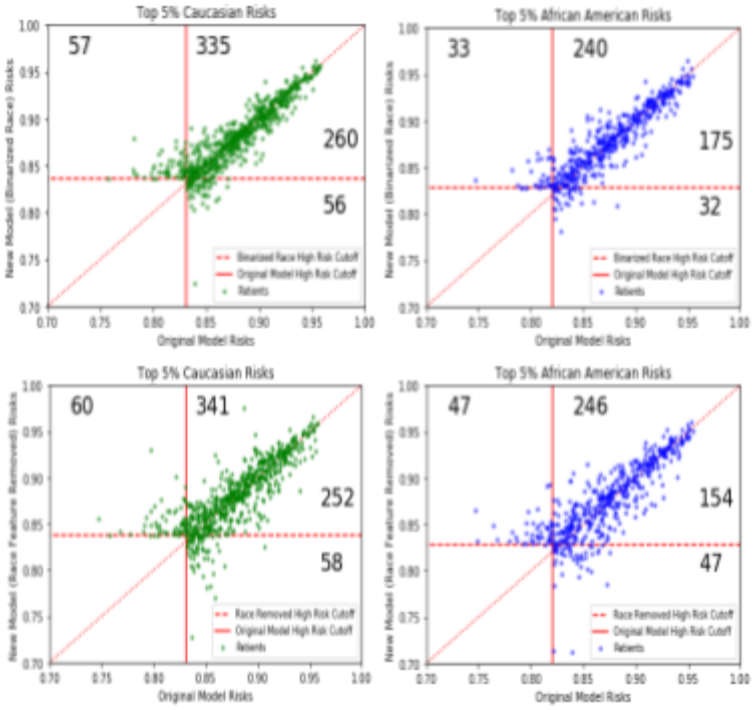
Background

Demographic race categorization in electronic health records (EHR) are often not self-reported, which has been shown to worsen discrepancies, especially in minority populations. These incorrect fields may influence machine learning model risk stratification. We tested the modification of race categorizations to look at its impact on a machine learning model that predicts inpatient hyperglycemic events.

Methods

38,269 patients from 52,317 inpatient encounters who received a high dose of corticosteroids (>=20 prednisone equivalents) at Duke University Hospital between Oct. 2014 and Aug. 2019 were used to train an XGBoost classifier. The outcome of interest was hyperglycemia (>=2 glucose values of 180mg/dL) within 24 hours of initial steroid administration. We trained three models that use race differently: 1.) the non-self-reported race categories from the EHR (original); 2.) binarized race labels (African American (AA) versus other) from the EHR; 3.) without race. 3773 features derived from 89 other clinical variables (e.g. vitals, laboratory values, medications, comorbidities) were used as inputs to the models. A 70%/30% train/test split was used for evaluation. After training the original model, we flipped categorizations from AA to Caucasian (C) and vice versa. We then revalidated on our test set.

Results



The plots visualize how the highest 5% risk cohorts vary among the three types of models. The x-axes show risks from the original models, and the y-axes in the top (bottom) row show risks from the binarized (no race) model. C (AA) high-risk populations are plotted in the left (right) column. The numbers on the graph areas denote counts in each region, and vertical and horizontal lines denote risk thresholds defining the highest 5% of scores. Numbers in the top-left of a subplot count patients who were in the highest 5% of risk for the new model but not the original model, while numbers in the bottom-right count patients who were originally high-risk but are not in the new model. The original model's AUROC was 0.874 (95% CI: 0.867 – 0.881), while the AUROC following flipped categorizations was 0.497 (95% CI: 0.475 – 0.518), making the performance no better than random. In AA categorized patients, only 82 of 455 patients (18%) remained in the top 5% risk cohort when their race was changed to C, and in C labeled patients, only 133 of 665 patients (20%) remained in the top 5% risk cohort when their race label was changed to AA.

Conclusion

Machine Learning for Healthcare 2020 – Clinical Abstract Track

Inaccurate capture of race information can negatively affect machine learning models, sometimes substantially changing predictions among the highest risk populations. It is unclear exactly what information recorded race fields capture. Future work is needed to assess how this information should be incorporated into machine learning models.