Giselle Miralles

Investigating racial discrepancies in MIMIC-III Database and Related Model's Labeling

I) Introduction

Machine learning algorithms have become a revolutionary aspect of modern medicine in recent years. With the ability to diagnose patients, predict mortality, disease outcome, and treatment outcome, these algorithms could be in the position to make life changing calls on a patient's health. Recent algorithms have proven the ability to detect the onset of sepsis in ICU patients' hours prior to onset. A public health record database, MIMIC-III, is a very popular choice to train machine learning algorithms, specifically in sepsis detection. This is due to its quantity of time-specific measurements on all aspects of patients' health.

With the deployment of such systems for critical decision making underway, it is extremely important that potential sources of bias are minimized across racial and ethnic groups. Several reports of machine learning algorithms deployed for risk-assessment in medicine display evidence of racial bias and exhibit a poorer performance in minorities. On top of this, many medical devices used to take vitals are shown to work less effectively for people with darker skin tones. Little work has been done to analyze the bias of these upcoming sepsis prediction models or the features in MIMIC-III used to train these models.

II) Background

The racial bias within machine learning algorithms deployed in the medical field have come under question. In 2019, [11], one study showed that an algorithm used by hospitals to determine the necessity of care in patients showed evidence of racial bias. This resulted in decreased detection for need of care for black patients than white patients. There are several examples of biased algorithms in the medical field, in both risk-assessment and diagnosis.

Many tools and existing measurement devices also show diminished performance in people of color [7]. Often this is due to a lack of representation in clinical trials and testing of these devices. These disparities can be even more magnified within the scope of AI in medicine. A recent study examined a model used to predict behavioral phenotypes using fMRI data [5]. It found differences in accuracy when trained on diverse and non-diverse data. With greater variance of the training data, they were able to improve the performance for racial minorities. Considering many potential sources of racial inequality in the creation of these models, it is important to identify and work to fix these in order to improve accuracy for all groups.

III) Data

The MIMIC-III dataset [6] is a very popular public database used to train machine learning algorithms for medicine. It includes data on patients staying in critical care units in the Beth Israel Deaconess Medical Center between 2001 and 2012. There is data on over 40,000 patients 16 years old and above. The data includes 26 tables that have information on patient demographics, lab test results, discharge summaries, diagnoses, and chart events [6]. Ethical considerations have been made; the database has undergone several de-identification measures including randomizing dates using date shifting, removing names, social security numbers, and addresses. Individual patient consent was waived due to these rigorous de-identification measures. One of the most popular applications of this database is to train machine learning models. The most common of these models is to predict patient outcomes in the ICU and predicting the onset of sepsis in the ICU.

Race/Ethnic Group	Number of admissions in the MIMIC-III Database
Native American/Pacific Islander	72 (0.12%)
Asian	2007 (3.4%)
Black	5785 (9.8%)
Hispanic/Latino	2145 (3.6%)
One or More Races	130 (0.22%)
White	41429 (70.24%)
Other	7408 (12.56%)

Table 1: Admissions by race/ethnicity in the MIMIC-III Database

III) Previous Models

Sepsis is the body's "overwhelming and life-threatening response to infection that can lead to tissue damage, organ failure, and death" [10]. Sepsis can quickly turn into septic shock, which is a severe illness with a very high mortality rate of 17-26% [9]. Current screening methods to determine if a patient is in fact suffering from sepsis can have high false positivity rates. Sepsis is also defined by a presence of several vague observations in patients, which is why it is so difficult to diagnose and treat promptly.

Many models so far have been created with the intention of detecting sepsis before it's critical stages. Because of the ample time stamped data in the MIMIC-III dataset, it is a very popular choice to train these models. In addition to training the model, labels must be made for each admission to identify the cases of sepsis through vitals and other variables. Some of the most prominent of these models include:

[1] Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach Desautels, Calvert, Hoffman, et al.	Insight model using elastic net regularization and 4-fold cross validation -AUROC (0.8799 [SD 0.0056])
[3] Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost Hou, Li, He, et al.	Conventional logistic regression model, SAPS-II score model and XGBoost algorithm model. -AUCs of 0.819 (95% CI 0.800–0.838), 0.797 (95% CI 0.781–0.813), and 0.857 (95% CI 0.839–0.876), respectively
[7] Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping Moor, Horn, Rieck, et al.	MGP-TCN and our DTW-KNN -AUC (0.7-0.9)

[8] MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis Rosnati, Fortuin	MGP-AttTCN -AUROC 0.6-0.69

Table 1: A selection of some of the most prominent studies using machine learning models used to detect sepsis onset, trained on the MIMIC-III database

Most used features among these models:

Heart rate
Respiration rate
SpO2
Temperature
Hemoglobin
GCS (Glasgow Coma score)
Glucose
Lactate
Age
Sodium
Anion Gap
Sodium Level
Systolic blood pressure
Diastolic blood pressure
White blood cell count
Potassium

Table 2: Most used features extracted from the MIMIC-III database from Table 1

IV) Methods

Access to the MIMIC-III Database through PhysioNet [6] is only granted to credentialed users. To access the dataset, training and user agreements were met in order to receive credentialed status. Once the dataset was loaded, it was processed and loaded into PostgreSQL. The dataset was processes using a script by Rosnati & Fortuin [5], replicating their process. Additional processing was done to standardize the racial groupings in the data and remove null/false values. Further analysis was then done in Python.

A) Feature investigation

Based on previous literature on the recurring and prominent features used in the above models, pulse oximetry and body temperature were selected for further investigation.

Pulse Oximetry:

Pulse oximeters measure the oxygen saturation in the blood, carried in your red blood cells. Pulse oximeters are typically clamped noninvasively to the fingertip for a quick and easy measurement of blood oxygen saturation. Because of the way pulse oximeters work, by sending light beams through the finger and recording the response, they are typically known to give more variable results in patients with darker skin tones [7]. This variability coupled with others have shown to contribute to bias ML algorithms deployed in the medical field [11].

In the MIMIC-III database, SpO2 level, which is recorded through pulse oximetry, is a recurring feature in the mentioned models. In order to understand its ethnic disparities in this dataset, it was compared with Arterial blood gas measurements in the dataset. While not a 1:1 similarity, SaO2 and SpO2 are typically correlated, and arterial blood gas is typically recorded through a physical blood sample, not exhibiting the same bias. The spread of SpO2 and SaO2 values matched on hospital admission ID is plotted.

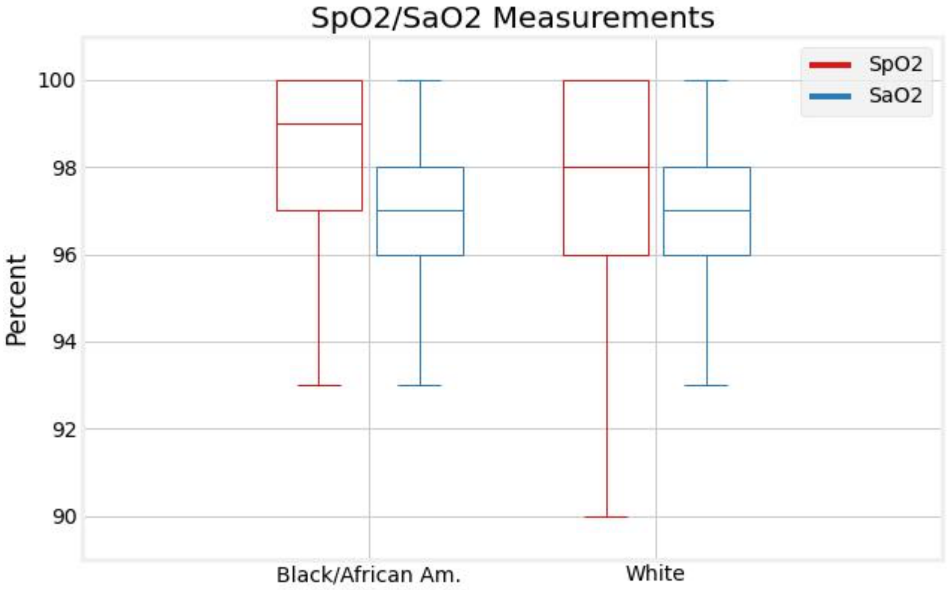

Figure 1: Plotted are the SpO2 measurements and paired SaO2 measurements for each hospital admission between self-identifying Black and White patients in MIMIC-III. Typical healthy measurements of both SpO2 and SaO2 are between 95% and 100%.

Race/Ethnic Group	Difference in means of SaO2 and SpO2 measurements (%)
Native American/Pacific Islander	-0.27
Asian	1.08
Black	1.57
Hispanic/Latino	1.24
One or More Races	1.25
White	0.81

Table 3: Differences in SpO2 and SaO2 measurements of each ethnic/racial group in MIMIC-III

Temperature:

For similar reasons as pulse oximeters, temporal thermometers have been shown to detect fever at lower rates for black patients versus white patients [3]. Their use of infrared radiation is hypothesized to interact differently with darker skin tones. General temperature measurements across patients were taken across ethnic groups.

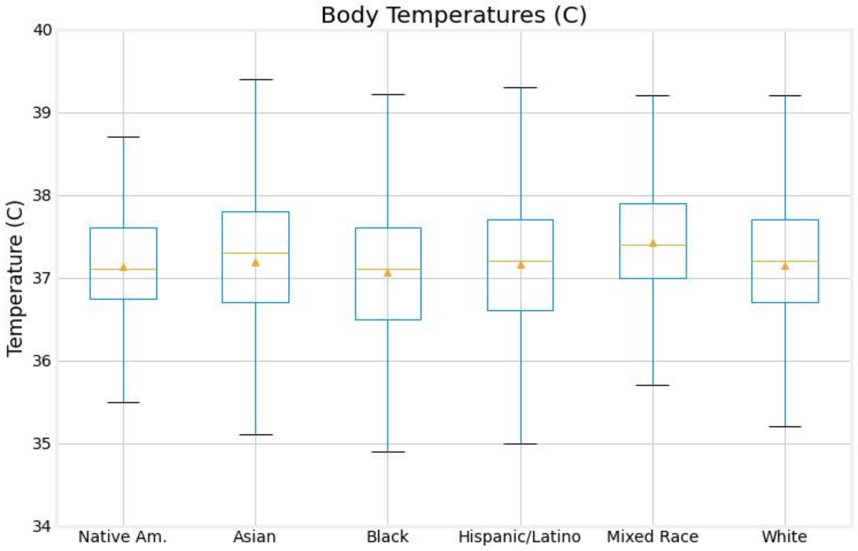

Figure 2: Differences in temperature measurements of each ethnic/racial group in MIMIC-III

B) Labeling methods from [9]

Rosnati's model to predict sepsis onset [9] claims to be an improvement on previous similar models achieving a ROC of 0.66. Rosnati's model, like others training on this dataset, required the production of labels on the patients to categorize them as developing sepsis or not within their admission to the hospital, since sepsis categorization can be ambiguous. The dataset was labeled based on predictors in the health statistics of MIMIC-III to determine whether and when a patient developed sepsis. A GitHub repository was provided for reproduction of these labels. Following this code and adjusting for bug fixes as necessary, the labels were reproduced.

Group	**Disp. Impact**	**% Labeled w/ Sepsis**
Native American/Pacific Islander	1.29	25%
Asian	0.74	14.6%
Black	0.90	17.6%
Hispanic/Latino	0.90	17.5%
One or More Races	1.15	22%
White	1.07	19%

Table 4: Racial breakup of patients labeled with sepsis among all patients in the MIMIC-III Database. Proportion of those labeled with sepsis within a racial grouping against the rest of the patients not in that grouping are recorded. Additionally, percent of each population labeled with sepsis is recorded

V) Results

When paired with Arterial Blood Gas measurements, pulse oximetry readings were shown to have a higher difference in readings in Black patients than White patients. Using Tukey HSD Multiple Comparison tests, significant differences are found between White/Black, White/Hispanic, White/Asian, Asian/Hispanic, Asian/Black SpO2-SaO2 levels. There are less significant racial differences in body temperature in the MIMIC-III dataset. This could be due to multiple types of thermometers being used without specification in the data's documentation.

Using M. Rosnati's labeling scheme, the spread of patients across ethnic groups is observed. Despite being reported having the highest sepsis mortality rates [1], under the labeling schema of Rosnati's predictive model, Black and Hispanic patients have among the lowest labeling rates as presenting with sepsis; Asians have the lowest rates, while Native Americans have the highest.

VIII) Conclusion

The MIMIC-III database reflects some of the existing racial discrepancies in the medical field. With the widespread use of this dataset for future and present medical breakthroughs in AI, it is important to be vigilant of the repercussions of this when training sepsis prediction algorithms. If used improperly, this could also worsen racial gaps in timely sepsis treatment.

Rosnati's Labeling scheme exhibits difference in the rate of classification of sepsis used to train their model. In conjunction with existing literature on the prevalence of sepsis in minorities, this divide may be even greater. Measures to lessen this bias within training and use of MIMIC-III should be taken into consideration by future models.

IX) Limitations and Future Work

While it would be ideal to also investigate results of existing studies, access to the predictions made by ML models was limited. Additionally, reproducing models with code/ documentation is difficult as even those that included GitHub repositories had errors, are outdated and include unfixable bugs. On top of this the large size of the data made it very difficult to work with on available technology. With respect to the contents of the database, there is not much documentation on specifically what is used for measurement. The MIMIC-III database is incredibly varied, and not completely consistent, therefore there are many variables that key for the same measurements and some assumptions had to be made as to which one exactly was used for models. Future work could involve investigating or replicating models trained on MIMIC-III to test for forms of racial bias and lokking into more features used in these models.

References

[1] Barnato, A. E., Alexander, S. L., Linde-Zwirble, W. T., & Angus, D. C. (2008). Racial variation in the incidence, care, and outcomes of severe sepsis: analysis of population, patient, and hospital characteristics. *American journal of respiratory and critical care medicine*, *177*(3), 279–284. https://doi.org/10.1164/rccm.200703-480OC

[2] Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., Shimabukuro, D., Chettipally, U., Feldman, M. D., Barton, C., Wales, D. J., & Das, R. (2016, September 30). *Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach*. JMIR medical informatics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5065680/

[3]*Emory researchers find temporal thermometers may miss fevers in black patients: Emory University: Atlanta ga*. alt-title. (n.d.). https://news.emory.edu/stories/2022/09/hs_bhavani_jama_racial_differences_thermometers_detecting_fevers_06-09-2022/story.html

[4] Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., & Wang, K. (2020, December 7). *Predicting 30-days mortality for mimic-III patients with sepsis-3: A machine learning approach using XGboost - Journal of Translational Medicine*. BioMed Central. https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-020-02620-5#Sec2

[5] Jingwei Li, et al., Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity.*Sci. Adv.***8**,eabj1812(2022).DOI:10.1126/sciadv.abj1812

[6] Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26.

[7] Nevarez, Freedman, McCullough, (2022, October 28). *Racial bias deeply rooted in healthcare technology and Medical Devices*. Racial Bias Deeply Rooted in Healthcare Technology and Medical Devices | The Chartis Group. https://www.chartis.com/insights/racial-bias-deeply-rooted-healthcare-technology-and-medical-devices#:~:text=In%20addition%2C%20research%20has%20found,populations%20prior%20to%20market%20launch.

[8] Moor, M., Horn, M., Rieck, B., Roqueiro, D., & Borgwardt, K. (2020, October 15). *Early recognition of sepsis with gaussian process temporal convolutional networks and dynamic time Warping*. arXiv.org. https://arxiv.org/abs/1902.01659

[9] Rosnati, M., & Fortuin, V. (n.d.). *MGP-ATTTCN: An interpretable machine learning model for the prediction of sepsis*. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0251248#abstract0

[10] *What is sepsis*. Sepsis Alliance. (2023, February 7). https://www.sepsis.org/sepsis-basics/what-is-sepsis/

[11] Ziad Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations.Science366,447-453(2019).DOI:10.1126/science.aax2342