Medical School Inequity Analysis*

Brent Maples[1], Max Marcum[2], and Randy Lin[3]

***Abstract*— This paper examines racial and ethnic disparities in medical school admissions using datasets from the Association of American Medical Colleges. By analyzing applicant and matriculant data from 2019-2023, including MCAT scores, GPA, and demographic information, the research highlights significant inequities faced by underrepresented groups, such as Black, Hispanic/Latino, and Native Hawaiian or Other Pacific Islander applicants. Logistic regression models reveal the impact of academic metrics and race on admissions, underscoring a lack of diversity in the future physician workforce. Through visualization and analysis, this work aims to inform actionable efforts to foster equity in medical education and address the gap between the diversity of the U.S. patient population and its healthcare providers.**

I. OBJECTIVE

In today's world, the role of doctors has become increasingly important. No longer do doctors solve the world's growing health needs, but they also research and explore to find new ways to cure others. According to the Center for Medicare and Medicaid Services [1], "U.S. health care spending grew 4.1 percent in 2022, reaching $4.5 trillion or $13,493 per person. As a share of the nation's Gross Domestic Product, health spending accounted for 17.3 percent." The importance of medicine today goes beyond the patient and has become a major player in the U.S. economy.

With such a big share in the economy, it is necessary that the major players in medicine, the doctors, represent patients from all backgrounds. Thus, it is the goal of this project to explore the racial inequities that exist in medical school admissions in order to inform readers on the potential barriers to entry that ethnic minorities face. Additionally, analysis of state information regarding ethnicities will be done to provide a comparison on states that are doing "well" and those at are not. Lastly, we intend to deliver this information via a web-based dashboard that will allow users to interact with and analyze the data whenever and wherever they want.

II. TERMINOLOGY

A. Terms

First, let us define the key identifiers used by AAMC[2] within their datasets.

- Applicant: A person who has applied to at least one U.S. MD-granting medical school through AMCAS and TMDSAS.
- Matriculant: A person who has applied to begin at an U.S.-MD granting medical school in a specific academic year and enroll in that academic year.
- State of Legal Residence: The self-reported state in which the applicant reports on their application that they reside.
- Race/Ethnicity: American Indian or Alaska Native; Asian; Black or African American; Hispanic or Latino; Middle Eastern or North African; Native Hawaiian or Pacific Islander; White; and Some other race or ethnicity.

III. APPROACH

We proposed three questions to be answered in this study:

- In this quickly expanding industry, what racial inequities are currently present in the medical school admission process?
- What are our doctors of tomorrow going to look like?
- Is there enough efforts in diversifying the doctor pool so that it accurately reflects the U.S. patient population?

IV. DATA

The datasets used were provided by the Association of American Medical Colleges (AAMC) [2] medical school admissions on their datasets page. The primary datasets chosen were those relating to race, ethnicity, and state information regarding applicants and matriculants. Due to the limitations present within this semester and the number of available group members, the study was conducted using data only from the 5-year period of 2019-2023. Below is a list of the AAMC datasets being used:

AAMC DATASETS

- Applicants to U.S. Medical Schools by Race/Ethnicity and State of Legal Residence
- Matriculants to U.S. Medical Schools by Race/Ethnicity and State of Legal Residence
- MCAT Scores and GPAs for Applicants and Matriculants to U.S. MD-Granting Medical Schools
- MCAT Scores and GPAs for Applicants and Matriculants to U.S. MD Granting Medical Schools by Race/Ethnicity
- MCAT Scores and GPAs for Applicants to U.S. MD-Granting Medical Schools by State of Legal Residence
- MCAT Scores and GPAs for Matriculants to U.S. MD-Granting Medical Schools by State of Legal Residence

The datasets studied above were primarily focused on the applicants and matriculants MCAT and GPA, with a greater emphasis on the "Total MCAT" score and none of the

*This work was not supported by any organization.

Brent Maples, Max Marcum, and Randy Lin are with the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville, USA. Emails: [1]bmaples6@vols.utk.edu, [2]mmarcu10@vols.utk.edu, [3]rlin8@vols.utk.edu.

MCAT subsections due to the MCAT score being cumulative. Applicants and matriculants were both studied because they represent the total applicant pool.

A. Processing the Data

Each dataset was split and merged so that applicants had their own separate datasets for each year and piece of information. This made it easier to process and display information such as applicants and their MCAT scores and GPAs for a state. The conversion process was done by transferring it to CSV from PDF and then using Python to make the data readable and usable for Python computations and JavaScript code.

V. Algorithms

In this project, a logistic regression model was used to identify and understand the potential connections between state, race, and MCAT/GPA. This was chosen because the relationship between all three can be summed into two categories:

- Accepted Into Medical School
- Declined From Medical School

Because these two options can be viewed as binary outcomes, it was chosen as the best method for studying the datasets. Within the codebase itself, the scikit-earn[3] implementation of the model was used due to its easy implementation.

Below is a pseudo-code implementation of the steps taken to use the applicant and matriculant data before incorporating it into the logistic regression analysis. The data and its features were then used to build the logistic regression analysis model and display important information regarding the outcomes.

Algorithm 1 Logistic Regression Implementation

Input: Data path, list of applicant files, list of matriculant files
Output: Model accuracy, classification report, logistic regression coefficients
Initialize an empty list `all_data` to store data for all years
for each pair of files (`app_file`, `mat_file`) in `applicants_files` and `matriculants_files` **do**
    Load applicants dataset from `data_path + app_file`
    Load matriculants dataset from `data_path + mat_file`
    Add column `Status = 0` to the applicants dataset
    Add column `Status = 1` to the matriculants dataset
    Combine applicants and matriculants datasets
    Append combined dataset to `all_data`
end for
Concatenate all datasets in `all_data` into a single dataset `data`
Define `target` as `'Status'`
Define `features`
Split `features`
Initialize logistic regression model
Train the model and Make Predictions
Evaluate the model:
- Compute and print accuracy score
- Compute and print classification report

Display logistic regression coefficients as:
- Feature name
- Corresponding coefficient value

VI. Primary Issues

A. Logistic Regression

Although the implementation of the logistic regression analysis worked, it came with its own skew of issues. Because the datasets each provided different features that could not be combined with each other, multiple regression models were implemented to be tested on each dataset type. Afterwards, a comparison of each coefficients impact were analyzed.

B. Data Conversion

Before cleaning the CSV datasets using Python, it first had to be transferred into a readable. Unfortunately, the AAMC datasets were all provided as PDFs rather than CSV files, thus it was very difficult to convert the PDFs into a digestible format for the code. In order to solve this, Convertio[4] was used to convert the datasets into a readable format, however it still made the data messy and required Python code to fix it completely.

VII. RESULTS

A. Logistic Regression with Race/Ethnicity

Using applicant race and ethnicity as features, the logistic regression model achieved a moderate accuracy of 75%. However, the model showed an imbalanced performance, as reflected in the lower recall for non-matriculants.

The coefficients in Table I revealed complex associations between race/ethnicity and matriculation. Positive coefficients for groups such as Black or African American ($+0.13$) and Hispanic/Latino ($+0.15$) suggest higher likelihoods of matriculation relative to other groups, potentially reflecting diversity efforts. Conversely, negative coefficients for groups like Native Hawaiian or Other Pacific Islander (-0.61) and Non-U.S. Citizens (-0.20) indicate challenges faced by these populations.

TABLE I
LOGISTIC REGRESSION COEFFICIENTS FOR RACE/ETHNICITY FEATURES

Feature	Coefficient
American Indian or Alaska Native	-0.083
Asian	0.137
Black or African American	0.126
Hispanic, Latino, or Spanish Origin	0.146
Native Hawaiian or Other Pacific Islander	-0.611
White	0.111
Other	0.148
Multiple Race/Ethnicity	0.091
Unknown Race/Ethnicity	0.021
Non-U.S. Citizen and Non-Permanent Resident	-0.201
Total	-0.115

The table above summarizes the average counts of applicants and matriculants by race/ethnicity, along with the differences between the two groups. It highlights disparities, such as a significant drop in the proportion of Black or African American and Hispanic/Latino applicants who matriculated, as well as an under-representation of Native Hawaiian or Other Pacific Islander populations.

TABLE II
APPLICANTS AND MATRICULANTS BY RACE/ETHNICITY

Race/Ethnicity	Applicants	Matriculants
American Indian or Alaska Native	1.71	0.69
Asian	241.15	113.35
Black or African American	89.58	35.37
Hispanic, Latino, or Spanish Origin	61.06	28.67
Native Hawaiian or Other Pacific Islander	1.06	0.37
White	405.92	183.29
Other	24.17	9.27
Multiple Race/Ethnicity	115.35	52.38
Unknown Race/Ethnicity	32.38	12.60
Non-U.S. Citizen and Non-Permanent Resident	13.10	3.04

The logistic regression analyses demonstrate that both academic metrics and racial/ethnic background play significant roles in medical school admissions. However, disparities persist, particularly among underrepresented groups, emphasizing the need for further efforts to promote equity and inclusivity in medical education.

TABLE III
DIFFERENCES BETWEEN APPLICANTS AND MATRICULANTS BY RACE/ETHNICITY

Race/Ethnicity	Difference
American Indian or Alaska Native	-1.02
Asian	-127.81
Black or African American	-54.21
Hispanic, Latino, or Spanish Origin	-32.38
Native Hawaiian or Other Pacific Islander	-0.69
White	-222.63
Other	-14.90
Multiple Race/Ethnicity	-62.96
Unknown Race/Ethnicity	-19.79
Non-U.S. Citizen and Non-Permanent Resident	-10.06

B. Graphing Results

Now, let us look at the population results over the 2019-2023 period generated by JavaScript graphs using the datasets. The logistic regression results noted that there was an intense disparity between Black or African American applicants, Hispanic/Latino applicants, and Native Hawaiian or Other Pacific Islanders in comparison to their Asian and White counterparts.

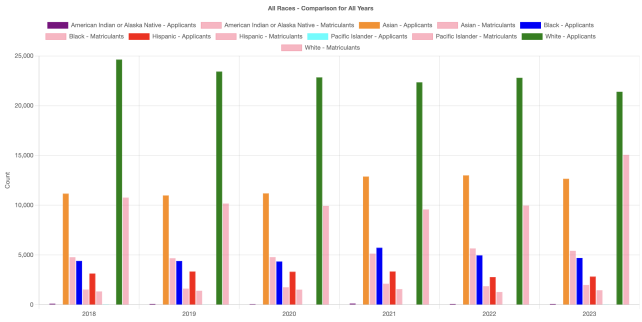

Fig. 1. Comparison of Applicants vs Matriculants by Race

As it can be seen from the figure above, there is a massive disparity between White applicants and other races. In fact, the second highest applicant rate, Asian, has applicant rates matching that of white matriculant rates. This indicates a severe problem in the USA's ability to encourage and foster future doctors that are representative of multiple races. Additionally, Black applicants and matriculants hold similar details to the White-to-Asian comparison, with Black applicants matching Asian matriculants numbers.

This is worrisome, as the Black population in the USA is 41.57 million while the Asian population is 19.12 million [5]. This indicates a massive disparity between minority involvement in medicine, which means that race representation in medicine is not representative of the USA race population.

Looking at the Black and Asian population here, it corroborates the results we have ascertained:

Now, let us take a closer look at the MCAT score breakdown that may be causing this: As it can be seen from the figure, Asian applicants and matriculants scored significantly higher than their Black counterparts. However, this data also shows that White's also scored less than Asians as well, which means that the number of applicants and potential

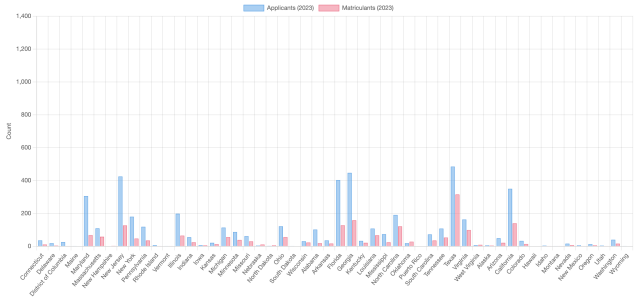

Fig. 2. Black Applicants vs Matriculants 2023

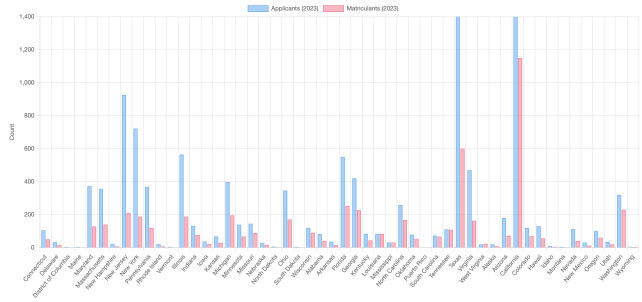

Fig. 3. Asian Applicants vs Matriculants 2023

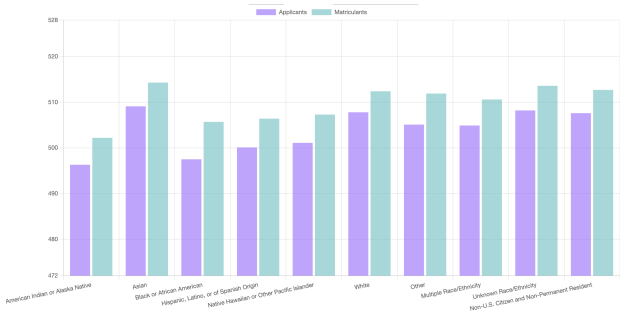

Fig. 4. 2023 MCAT Scores by Race

racial biases may lead to higher acceptance rates for Whites. Additionally, this data shows that there may not be enough efforts to encourage and educate minority applicants to become more competitive in the medical schools admission process.

VIII. CONCLUSION

In the beginning, we proposed three questions and have found that this data has helped answer all of those questions. In response to: "In this quickly expanding industry, what racial inequities are currently present in the medical school admission process?", we find that there are significant racial inequities present that are not encouraging minority applicants such as Black and Hispanic applicants to contest with the majority of White and Asian applicants. Additionally, we find that those with the lowest level of admissions also have the lowest MCAT scores. This means that there must be a greater initiative to encourage, help, and educate minority applicants to go to medical school and help them succeed.

Thus, this answers our next question: "What are our doctors of tomorrow going to look like?". In fact, it seems that the doctors of tomorrow are going to represent a part of population, but still do not represent the diverse group of people within the USA. The lacking number of Black applicants in comparison to their 41.57 million population.

Lastly, we asked ourselves: "Is there enough efforts in diversifying the doctor pool so that it accurately reflects the U.S. patient population?" Although this data struggles to answer this fully, it does show a growing divide between doctors and their diverse patient population. While the USA population has steadily increased over the years, we must ask ourselves if skilled labor such as medical doctors are in-pace with that as well.

IX. FUTURE WORK

This brings us to our future work, which we believe is incredibly important to analyzing the full future of making medicine equitable. We now aim to find data and study questions that look at local communities within a state to analyze what can be done there to answer the question: "Is there enough efforts in diversifying the doctor pool so that it accurately reflects the U.S. patient population?" In this, we will reach out to medical schools and hopefully obtain data that will be more useful in answering these questions. Additionally, we will reach out to local communities and states to learn and seek out the initiatives that are encouraging minorities to become involved in medical school.

X. OTHER

Below is important information regarding the teams purpose, timeline, and responsibilties.

A. RESPONSIBILITIES

Our team is composed of three members, all of them Computer Scientists.

- **Brent Maples**: Lead project, research datasets, write research paper, build front-end/back-end code-base.
- **Max Marcum**: Research lead, write research paper, investigate machine learning, build back-end code-base
- **Randy Lin**: UX/UI lead, write research paper, research datasets, investigate UI, build front-end code-base.

B. TIMELINE

Below is a timeline of the work that was done and the deadlines that came with it. This project did not fall behind at any point.

- **September 27th, 2024** - Conversion of all datasets to readable format
- **October 15th, 2024** - Initial GUI Creation
- **October 27th, 2024** - Initial Creation of Data Analysis Back-end
- **November 1st, 2024** - Initial Creation of Front-End
- **November 4th, 2024** - Begin Paper
- **November 6th, 2024** - Deliver on Minimum Viable Product (MVP)
- **November 18th, 2024** - Begin Presentation

- **November 20th, 2024** - Finalize Front-End/Back-End
- **November 22nd, 2024** - Finish Writing of Paper
- **November 27th, 2024** - Finish Presentation
- **December, 2024** - Present Project

C. OUTCOME

The intention of this paper is to provide an easy way to inform institutions of the potential racial inequities in medical school admissions. The paper does not dictate what actions should be taken, but rather leave it up to the reader to infer what steps can be taken next. The information available is a helpful resource in educating in-state medical schools and encouraging institutions to potentially improve their racial equity in the medical schools admission process.

REFERENCES

[1] C. for Medicare Medicaid Services, "National health expenditure data: Historical," https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/historical, 2023, accessed: 2024-09-10.

[2] Association of American Medical Colleges, https://www.aamc.org/, 2023, accessed: 2024-09-10.

[3] S. learn Developers, "Logistic regression — scikit-learn 1.5.0 documentation," 2024, accessed: 2024-11-18. [Online]. Available: https://scikit-learn.org

[4] Convertio, "Pdf to csv converter — convertio," 2024, accessed: 2024-11-18. [Online]. Available: https://convertio.co/pdf-csv/

[5] United States Census Bureau, "American community survey 5-year data (2009-2021)," 2024, accessed: 2024-11-18. [Online]. Available: https://www.census.gov/data/developers/data-sets/acs-5year.html