BIOS 511 Final Project Report

Bryan Jacobs

Introduction and Literature Review

The prevalence of slavery in the United States from the early 17th century defines the darker side of America's history. Even after events like the ratification of the 13th Amendment and the Civil Rights Movement, structural inequities had already been formed between demographic groups. These inequities, often examined via critical race theory today, still largely affect many citizens of the United States.

One of the sectors currently exhibiting the greatest effects of historical racial inequities is the United States healthcare system. Decades of research have demonstrated this, with one recent academic paper claiming, "An overwhelming body of evidence points to an inextricable link between race and health disparities in the United States" (Macias‑Konstantopoulos et al., 2023). This paper, published in the National Library of Medicine, goes on to list a multitude of diseases and complications in which minority groups tend to be negatively impacted to a greater degree than their white-identifying peers. For example, "In 2013, Blacks and Hispanics accounted for 46% and 21% of new HIV infections and 49% and 20% of new AIDS diagnoses despite representing 12% and 16% of the total US population, respectively" (Macias‑Konstantopoulos et al., 2023). When discussing each complication, the authors outline actionable steps to be taken in order to work towards reducing these disparities. By far, the most common actionable step listed is increasing access to care. This raises the question: Do government insurance programs, such as Medicaid and Medicare, reduce disparities in healthcare system utilization across demographic groups?

Data Source and Preparation

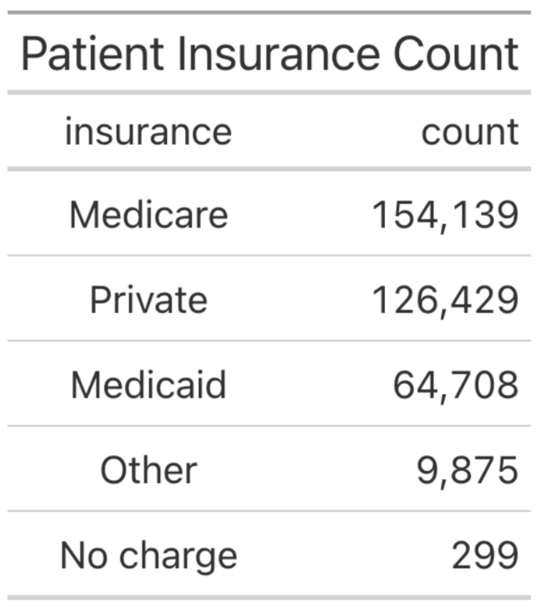
The data used in this analysis comes from the MIMIC-IV public EHR dataset. This dataset was accessed via an academic referral along with completion of a human research subject protections and HIPAA ethics course. The specific datasets used were the *admissions* and *patients* datasets. Together, these datasets contained information about 546,028 unique hospitalizations of 223,452 unique individuals.

With such a large amount of data, the best method for dealing with NA values was to omit records containing them in important features. After selecting only the necessary columns, the data was coerced in such a way that each record represented one patient in one year, with the response variable defined as the number of hospital visits per year per patient. This method produces a count dataset, which is important for downstream analyses. The main predictors were insurance type and race/ethnicity, while age and marital status were used to control for confounding. For interpretability, insurance type and race/ethnicity were entered into the model as an interaction term. Because of the nature of count data, only patient-level variables were possible to include. Important, potentially confounding hospitalization-level variables such as admission type and location could not be included, a critical limitation to consider when interpreting results.

Data Analysis and Results

Exploratory Data Analysis

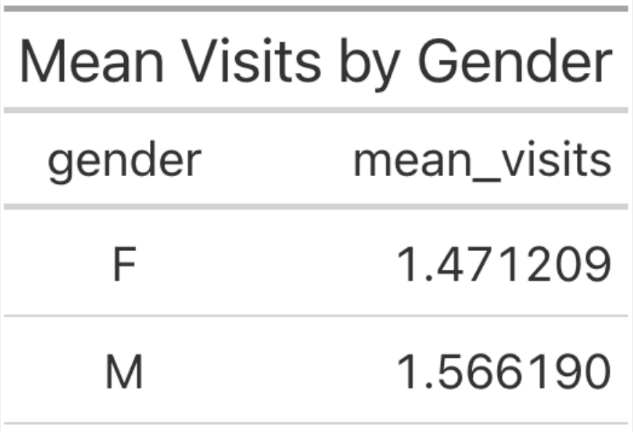
The first necessary exploratory data analysis was a verification of data density and balance.



Patient Insurance Count

insurance	count
Medicare	154,139
Private	126,429
Medicaid	64,708
Other	9,875
No charge	299

Fig. 1



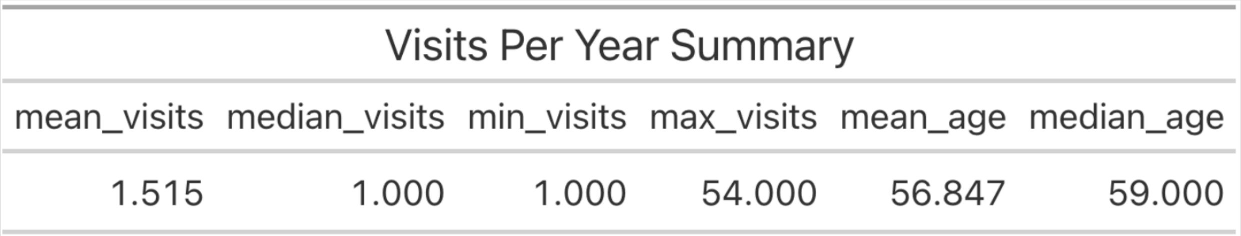
Mean Visits by Gender

gender	mean_visits
F	1.471209
M	1.566190

Fig. 2

Figure 1 shows a sufficient number of data points in the insurance categories of interest: Medicare, Private, and Medicaid. *Figure 2* shows the data is relatively balanced between mean hospital visits per year by men and women, each averaging about 1.5 visits per year. These metrics suggest a strong foundation for downstream analysis.

Next, the distribution of the response variable, hospital visits per year per patient, was assessed.



Visits Per Year Summary

mean_visits	median_visits	min_visits	max_visits	mean_age	median_age
1.515	1.000	1.000	54.000	56.847	59.000

Fig. 3

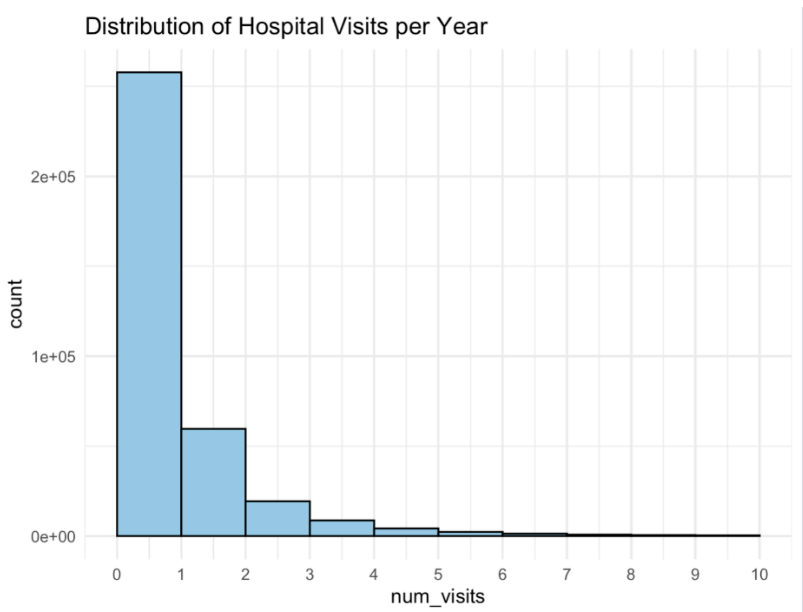
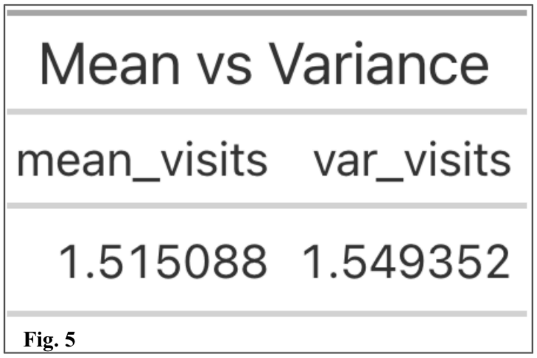


Fig. 4

Ideally, the response variable would be normally distributed. However, *Figures 3* and *4* show a right skew, with roughly 70% of the datapoints having 1 visit per year. This fact is critical to consider when conducting downstream analysis.

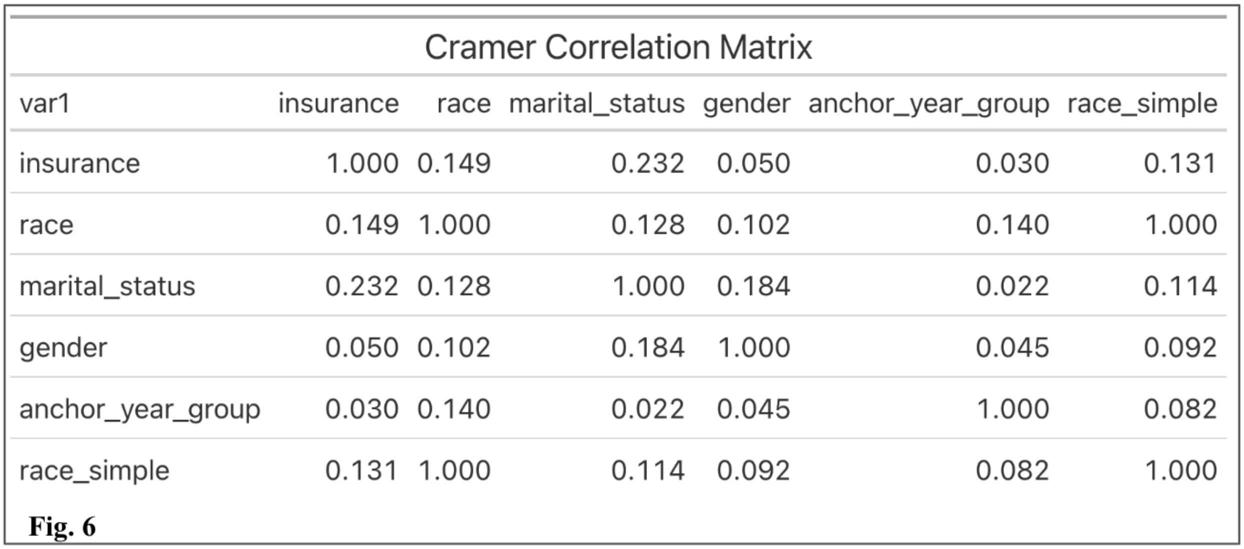
Additional exploratory analyses examined the mean versus variance of the response variable and the correlation between model variables.



Mean vs Variance

mean_visits	var_visits
1.515088	1.549352

Fig. 5



Cramer Correlation Matrix

var1	insurance	race	marital_status	gender	anchor_year_group	race_simple
insurance	1.000	0.149	0.232	0.050	0.030	0.131
race	0.149	1.000	0.128	0.102	0.140	1.000
marital_status	0.232	0.128	1.000	0.184	0.022	0.114
gender	0.050	0.102	0.184	1.000	0.045	0.092
anchor_year_group	0.030	0.140	0.022	0.045	1.000	0.082
race_simple	0.131	1.000	0.114	0.092	0.082	1.000

Fig. 6

When modeling count data, a Poisson regression is often the first step. For a Poisson regression to be appropriate, the mean and the variance of the response variable must be roughly equal. With a mean of about 1.52 and a variance of about 1.55 shown in *Figure 5*, the equality is ambiguous, and final model selection must rely on other metrics. All potential models for this count dataset require independence. *Figure 6* shows a Cramer Correlation Matrix of the model variables, with the greatest correlation of 0.232 appearing between insurance and marital status. With correlations of less than or equal to 0.232, multicollinearity is not likely to pose a problem.

Methods

As previously mentioned, the first option for modeling count data is a Poisson regression, but this is only viable when the mean and variance of the response variable are roughly equal. Because the mean and variance here are close but not conclusively equal, it was unclear whether the Poisson regression would be overdispersed. Therefore, three separate regression models were fit and compared: a Poisson regression model, a negative binomial regression model (more robust to overdispersion), and a quasi-Poisson model (more robust to both under- and overdispersion).

After training the models on the same 70% of the data and testing on the remaining 30%, the Poisson model showed no overdispersion. Because the other models use the same underlying linear predictor, they produced identical results. Therefore, we proceeded with the Poisson model.

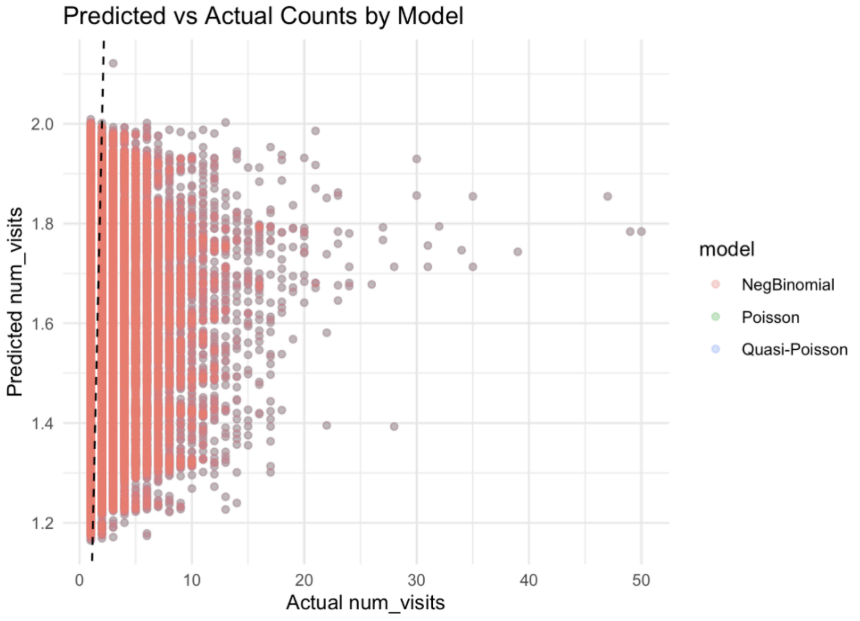
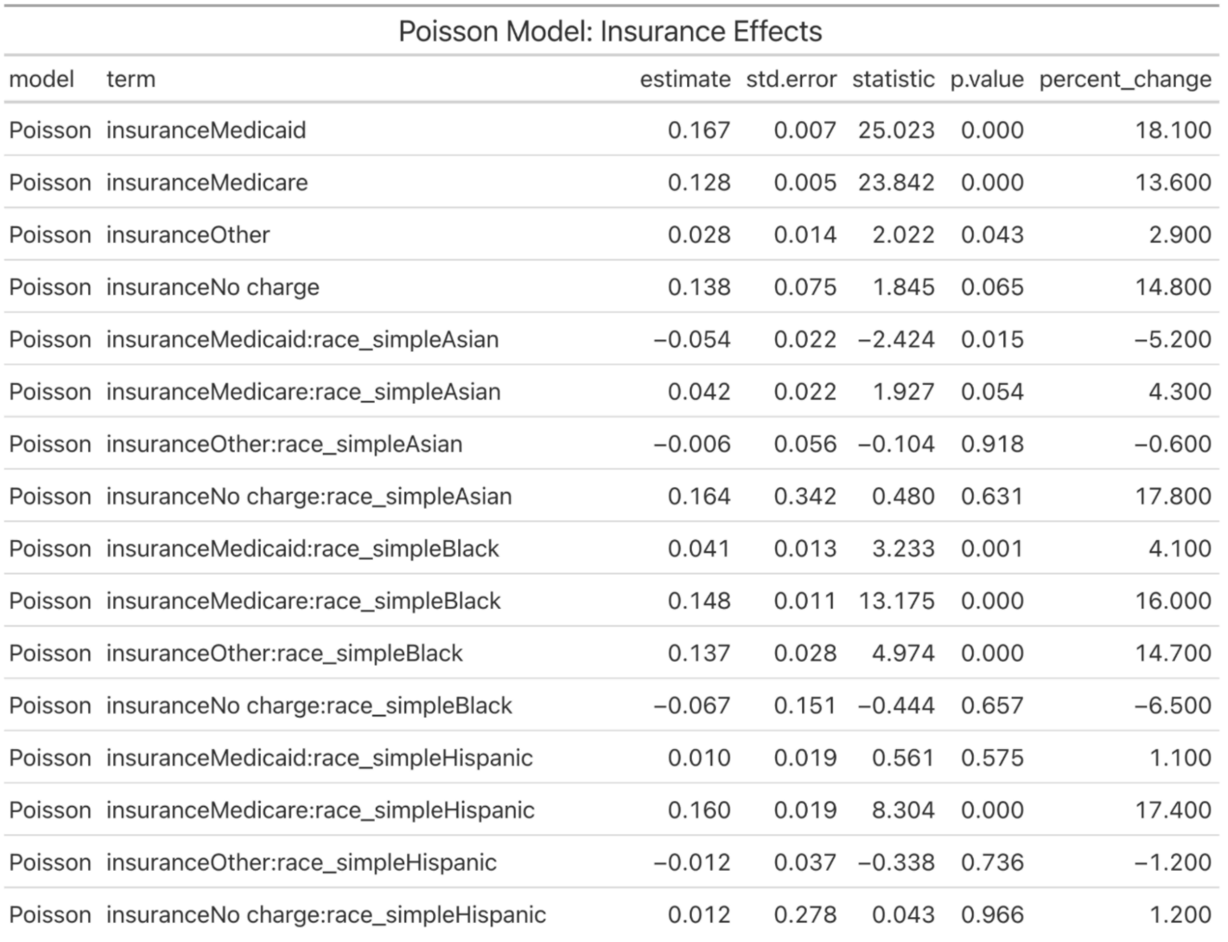


Fig. 7

The model exhibited a root mean squared error (RMSE) of 1.24. Considering the possible number of visits in the dataset, this seems reasonable, but given the fact that 70% of data points showed only one visit per year, this value is not ideal. As shown in *Figure 7*, this relatively high RMSE comes from the concentration of data at one visit per year pulling almost all predictions, regardless of actual value, below two.

Results



Poisson Model: Insurance Effects

model	term	estimate	std.error	statistic	p.value	percent_change
Poisson	insuranceMedicaid	0.167	0.007	25.023	0.000	18.100
Poisson	insuranceMedicare	0.128	0.005	23.842	0.000	13.600
Poisson	insuranceOther	0.028	0.014	2.022	0.043	2.900
Poisson	insuranceNo charge	0.138	0.075	1.845	0.065	14.800
Poisson	insuranceMedicaid:race_simpleAsian	−0.054	0.022	−2.424	0.015	−5.200
Poisson	insuranceMedicare:race_simpleAsian	0.042	0.022	1.927	0.054	4.300
Poisson	insuranceOther:race_simpleAsian	−0.006	0.056	−0.104	0.918	−0.600
Poisson	insuranceNo charge:race_simpleAsian	0.164	0.342	0.480	0.631	17.800
Poisson	insuranceMedicaid:race_simpleBlack	0.041	0.013	3.233	0.001	4.100
Poisson	insuranceMedicare:race_simpleBlack	0.148	0.011	13.175	0.000	16.000
Poisson	insuranceOther:race_simpleBlack	0.137	0.028	4.974	0.000	14.700
Poisson	insuranceNo charge:race_simpleBlack	−0.067	0.151	−0.444	0.657	−6.500
Poisson	insuranceMedicaid:race_simpleHispanic	0.010	0.019	0.561	0.575	1.100
Poisson	insuranceMedicare:race_simpleHispanic	0.160	0.019	8.304	0.000	17.400
Poisson	insuranceOther:race_simpleHispanic	−0.012	0.037	−0.338	0.736	−1.200
Poisson	insuranceNo charge:race_simpleHispanic	0.012	0.278	0.043	0.966	1.200

Fig. 8

Figure 8 shows the results for the insurance and insurance × race interaction terms. Percent changes were calculated for interpretability. While p-values are not the most informative statistic in Poisson regression modeling, it is still notable that many of them are relatively small. The most important takeaway is that most percent changes are significantly positive. In an ideal world, a percent change of 0 would indicate no disparities. However, given the historical inequities in the healthcare system, positive percent changes in insurance × race interaction terms, relative to private insurance × white race, indicate progress in reducing disparities in access to care.

Positive percent changes on insurance features suggest that government insurance programs are associated with increased use of the hospital system overall. Positive percent changes on the interaction features suggest increased hospital usage among minority groups, with a greater effect in older populations.

Beyond these statistical conclusions, it is difficult to draw a definitive inferred conclusion. One possible interpretation, our motivating hypothesis, is that government insurance programs reduce disparities in access to care. Another equally plausible interpretation is that historical inequities have led to increased hospitalization among minority patients. Determining the more accurate conclusion would require interdisciplinary research involving sociologists, anthropologists, policymakers, and other relevant experts.

Limitations and Conclusion

Limitations

Research on a complex social topic like this inevitably comes with limitations. One major limitation was selection bias due to structural inequities. The historical and systemic inequities previously discussed influence who uses the American hospital system and, therefore, who appears in the dataset. The study disregards individuals who do not access the hospital system at all, ironically, the very individuals whom policies aim to reach.

Confounding is another limitation. Dataset and count-data constraints prevented inclusion of important hospitalization-level variables. Future studies would benefit from incorporating variables such as health status, comorbidities, and transportation access. Although many of these factors are themselves structurally linked to race, controlling only for race is insufficient.

Measurement limitations also pose challenges. Hospital and emergency department visits were used as a proxy for access to care, but this excludes primary care visits, an essential part of the healthcare landscape. High emergency department use is often indicative of poor primary care access.

Finally, it is important to recognize that race is used as a proxy throughout this study. Race is a social construct, not a biological determinant. Observed disparities therefore reflect structural inequities, not inherent biological differences.

Future Improvements

To computationally improve the study, the model must better account for the concentration of observations at one visit per year. A zero-inflated or hurdle model is typically used when the response variable contains many zeroes and could be adapted for this scenario where the modal value is one. Additionally, including populations that do not use the healthcare system would reduce selection bias. Incorporating hospitalization-level confounders and adding primary care visit data would also strengthen the analysis.

Conclusion

While the results of the study alone are insufficient to draw concrete conclusions, they represent a step toward evaluating whether government insurance programs reduce disparities in healthcare system utilization across demographic groups. With appropriate computational refinements and interdisciplinary collaboration, this line of research has the potential to guide policymakers and assist in making informed decisions aimed at addressing historical inequities.

References

Macias-Konstantopoulos, W. L., Collins, K. A., Diaz, R., Duber, H. C., Edwards, C. D., Hsu, A. P., Ranney, M. L., Riviello, R. J., Wettstein, Z. S., & Sachs, C. J. (2023). Race, healthcare, and health disparities: A critical review of emergency medicine. *[Article]*. *PMC free article*.