Examining Unhoused Hospitalizations and Emergency Department Visits in California to Create Policy Recommendations

Govind M. Cleckler

1. INTRODUCTION

As homelessness has continued to explode across the United States since the mid 1980's, California has seen itself at the center of the issue, with 28% of the country's unhoused population residing there. The fact that 51% of all unsheltered people in the country were in California as of 2020 shows the pressing need to both ensure safety measures are effective and medical care is accessible for those that require it. This study was guided by interest in understanding the capabilities of the health care systems in California and the impact that bias and underfunding has on them. This paper offers policy-driven recommendations based upon analysis of hospital encounters for both unhoused and housed patients in California hospitals. This study uses social identifiers to determine the impacts and effectiveness of different policies and laws within various communities and demographics.

2. DATA

2.1. Background

2.1.1. Data Dictionary

Below are selected descriptions of important variables and their descriptions from the data set.

Variable	Description
EncounterType	Inpatient (IP) Hospitalization or Emergency Department (ED) Visit
HospitalCounty	The county where the treating hospital is located
FacilityName	The name of the treating hospital
System	The name of common ownership and/or association for a group of facilities.
Ownership	The hospital's ownership type (Government, Investor, or Non-Profit)
Urban_Rural	Indicates if hospital has an Urban or Rural (includes Rural/Frontier) designation
Teaching	Indicates if the hospital is designated as a teaching hospital
LicensedBedSize	The hospital's number of licensed beds
PrimaryCareShortageArea	Indicates if the hospital is located in a Shortage Area for Primary Care
MentalHealthShortageArea	Indicates if the hospital is located in a Shortage Area for Mental Health Care
HomelessIndicator	Indicates if the data is for Homeless or Non-Homeless encounters
Demographic	Age, Race, Sex, or (Expected) Payer. Other Payer includes Workers' Compensation, Other Government, Title V, Disability, VA Plan, Other Payer, invalid, and missing
DemographicValue	Value for demographic category
Encounters	Count of inpatient hospitalizations or emergency department visits
TotalEncounters	Total inpatient hospitalizations or emergency department visits per hospital.
Percent	Calculation: Encounters/Total Hospital Encounters x 100

2.1.2. Initial Characteristics

The 2019-2020 Homeless Hospital Encounters: Age, Race, Sex, Expected Payer By Facility dataset[1], published by California's Department of Health Care Access and Information, contains the data for inpatient hospitalizations and emergency department (ED) visits for both housed and unhoused people. This report is the most recent that is available and originally contained 432,633 data points. The dataset includes hospitalization counts at various locations based upon a number of factors, including housing status (HomelessIndicator), Demographics (Age, Sex, Type of Insurance, Race), and Demographic Value, which provides the data for each Demographic row. However, this method of organization resulted in duplicate data points in the Encounters column, making any meaningful analysis impossible without a restructuring of the dataset. All variables were initially categorical, besides Encounters and Total Encounters, which are numerical.

2.1.3. Additional Features

Variables that stuck out as possible areas of interest for analysis include, Ownership, Urban_Rural, Teaching, LicensedBedSize, PrimaryCareShortageArea, and MentalHealthShortageArea. All of the values in these columns were written as strings, instead of integers, which would have been much simpler to analyze. Since these columns all contain information about the hospitals themselves rather than unhoused patients, they function more effectively as data for examining those hospitals.

2.2 Data Cleaning

The dataset required cleaning to get usable numbers in order to conduct correlations between different variables, such as demographic information and number of hospital beds. To clean the dataframe, a separate dataframe named *results* was created using the pivot_table() function in pandas. This allowed the duplicate values in the Demographic column to be condensed into one row for inpatient hospitalizations and one for emergency department visits. A new dataframe, dff, was created to place the updated values and drop the old Demographic, DemographicValue, TotalEncounters, Encounters, and Percent columns. Columns that had true/false values were transformed to 0s and 1s using the pd.get_dummies() function. This increased the number of variables that are useful for finding trends. This transformation also let columns that were previously strings become integers so that a machine learning model could be applied. The final dataframe, *finaldf*, contains 1,497 rows and 29 columns, for a total 43,413 unique data points.

[1] 2019-2020 Homeless Hospital Encounters: Age, Race, Sex, Expected Payer By Facility | CA Open Data. https://lab.data.ca.gov/dataset/hospital-encounters-for-homeless-patients/8e61fb2e-2dc7-4fb7-a6bf-250345976dae. Accessed 26 May 2024.

3. EXPLORATORY DATA ANALYSIS

3.1. Research Question

This portion of the research was guided by the question: how do social identifiers impact the unhoused hospital care? The demographic variable immediately stood out as interesting and an area for possible research because determining which identifiers are predictors for a high risk person can create a clearer picture of certain failures or victories of the healthcare system.

3.2. Graphs

Once the data cleaning phase was completed, the data analysis began with a correlational heatmap using the cleaned *finaldf* dataset. As seen below in Figure 1, the dataframe contains aggregate counts for each hospital's cases for different social demographics, meaning that correlations from the Sex_Male row to the Sex_Male column contain the most valuable insights.

3.2.1. Demographic Heatmap

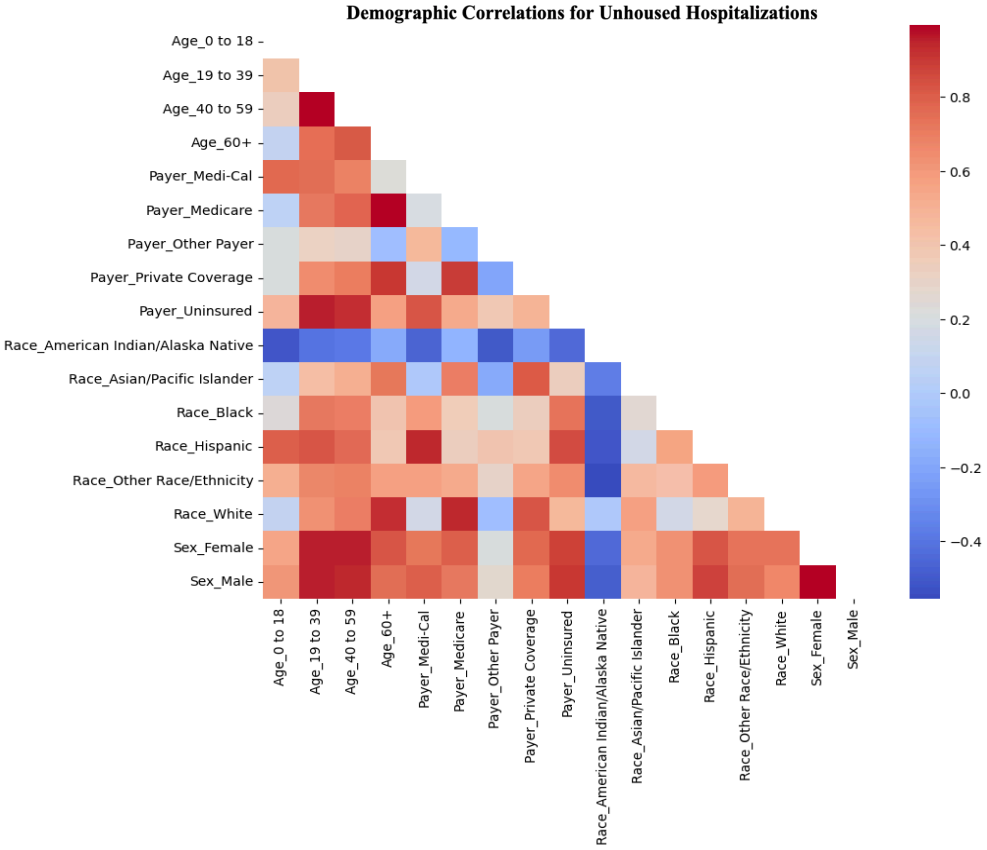

Figure 1. Values to the right of the diagonal were removed due to repeated correlation values that made the graph more visually confusing. Using the *pandas* and *numpy* libraries in python, correlations between 0.5 and -0.5 were not considered to filter unuseful information.

With values where conclusions could be determined with confidence (those in the Demographic column), some of the most striking positive correlations were:

- *Medicare & Age 60+: Being over 60 is a reliable predictor of being on medicare*
- *Hispanic and Medi-Cal: Being Hispanic is reliable for predicting Medi-Cal*
- *Uninsured and Age 19-39: People from ages 19-39 correlates with not having insurance*

Since all of these features come from inpatient and emergency department visits, it shows that these demographics, other than Medicare & Age 60+,[2] all have a disproportionately high risk of hospitalizations.

The most striking negative correlations were between:

- *Other Payer & White*
- *Other Payer & Age 60+*

Other observations:

- Values in the American Indian/Alaska Native mostly contain sample sizes < 10 incidents per facility, so those negative correlations are less noteworthy.
- Unhoused hospitalization rates for people over 60 and White people are disproportionately low.[3]

These negative correlations demonstrate that groups in the Other Payer category (such as workers compensation, government, or disability) are at greater risk of unhoused hospitalization and are in need of additional support.

3.1.2 Income Comparison

Using a simple, three-column dataframe from the National Institute on Minority Health and Health Disparities, titled *Income (Median household income) for California by County,* a box plot (Figure 2) of household income was compiled by county in order to visualize average income in California. Prior to the data cleaning step, the dataframe contained a

[2]"Homelessness and Racial Disparities." National Alliance to End Homelessness, https://endhomelessness.org/homelessness-in-america/what-causes-homelessness/inequality/. Accessed 14 June 2024.
[3] Ibid.

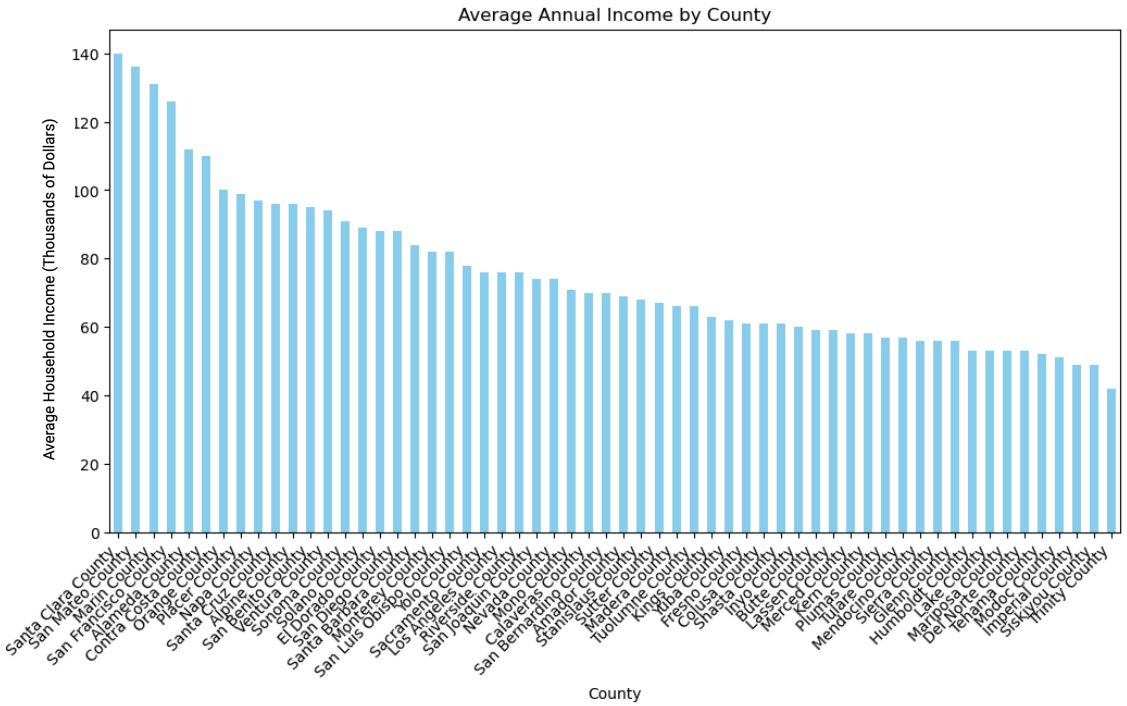

Figure 2

column entitled ‘HospitalCounty,’ which allowed comparisons between household income for both unhoused and housed hospitalizations. This was ultimately used to examine policy decisions and try to identify factors in these counties that cause discrepancies or outliers. This dataset is used in *Figure 3* to help determine how income impacts the capabilities of a hospital.

3.1.3 Hospital Capabilities (LicensedBedSize) & Average Income

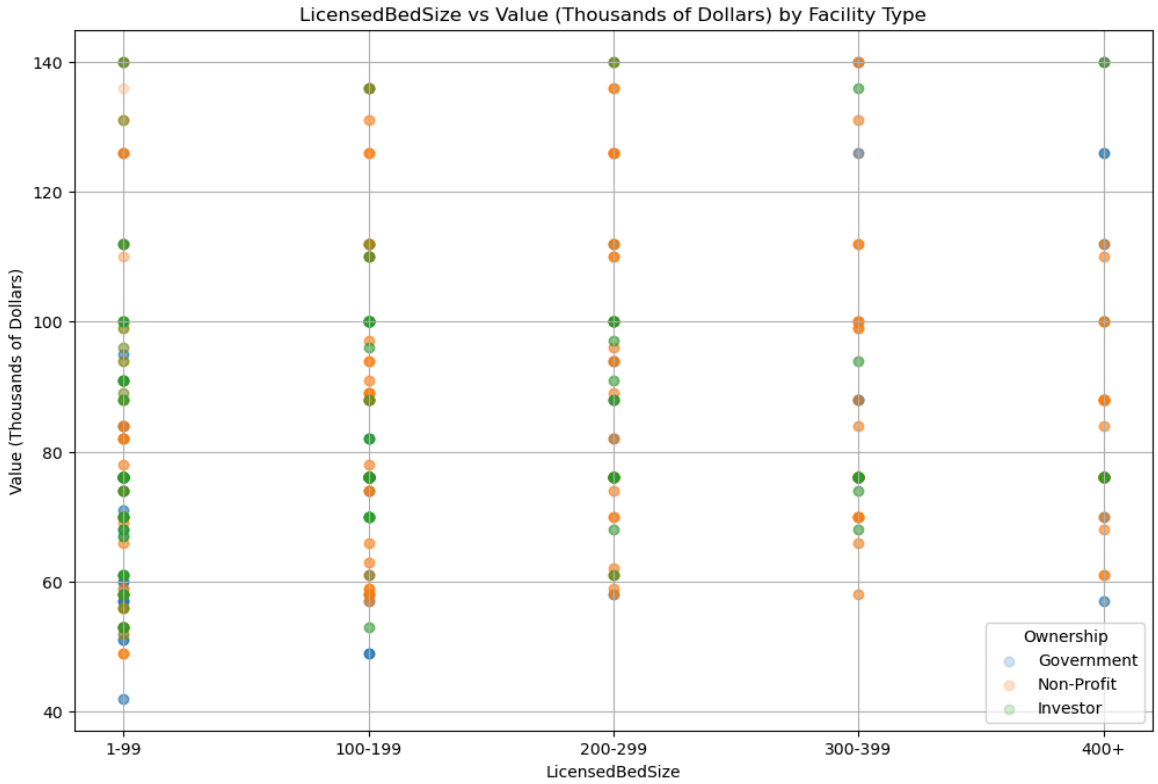

Figure 3. This figure shows the number of available beds in a given hospital plotted by the average household income of the county that the hospital is in. Also included in the color of the data points is information about who owns each hospital (government, non-profit, or investor).

This visualization demonstrates that many of the hospitals with a low number of beds are investor-owned in counties earning less than $100,000/year. Chain investor-owned hospitals are much larger than singular hospitals in terms of the number of beds,[4] which doesn't match with the 1-99 beds column in the figure. The hospitals that are best prepared are most commonly non-profit, with strong representation with 300+ beds, regardless of income. It should be noted that some hospitals may not require high bed counts to adequately serve their communities.

4. MACHINE LEARNING/MODELING

4.1 Benefits of Machine Learning

The benefit of creating a machine learning model for this dataset lies in the ability to find correlations. 70% of the known values for the dataset were used to train the model, and the remaining 30% were used to test the model's effectiveness. By

[4] Medicine (US), Institute of, and Bradford H. Gray. "Legal Differences Between Investor-Owned and Nonprofit Health Care Institutions." The New Health Care for Profit: Doctors and Hospitals in a Competitive Environment, National Academies Press (US), 1983. www.ncbi.nlm.nih.gov, https://www.ncbi.nlm.nih.gov/books/NBK216759/.

comparing the two correlations, an accuracy rating between 0 and 1 is created to ensure that the model is accurate enough to draw conclusions from. The model used was a random forest regression model, which is an ensemble technique that combines a multitude of decision trees into one large "forest" model.[5] The result is that each feature has a number of importance that tells us which identifiers demonstrate the highest correlation, and therefore which areas are most important to focus on when trying to address issues such as primary or mental health care shortage. The results included all have an importance level greater than 0.05 since anything lower is not reliable for prediction.

4.2 Primary Care Shortage Areas

We created a model that uses unhoused inpatient care data to predict primary care shortage areas which gave an acceptable accuracy of 78.3%. Average household income had the strongest positive correlation, along with being in an area with a shortage of primary care; lower-income areas often have underfunded health facilities.[6] Since this correlation was already well-established, it was removed from the training data in order to assess factors that were less apparent. The two next strongest predictors, living in an urban area or being Hispanic, offered insight into which populations are most underserved by areas that lack primary care access. These correlations show yet another aspect of the effects of underfunding in redlined areas in California.[7]

Predictors	Feature Importance (MDI[8])
Urban	0.158990
Hispanic	0.071174
Black	0.062748
Private Insurance	0.062185
White	0.050349

4.3 Mental Healthcare Shortage Areas

The model also used unhoused inpatient care data to predict mental healthcare shortage areas, giving an acceptable accuracy of 72.9%. Average household

[5] "A meta estimator that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting." https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#:~:text=A%20random%20forest%20regressor.,accuracy%20and%20control%20over%2Dfitting.
[6] Hussein, Mustafa, et al. "Neighborhood Socioeconomic Status and Primary Health Care: Usual Points of Access and Temporal Trends in a Major US Urban Area." Journal of Urban Health : Bulletin of the New York Academy of Medicine, vol. 93, no. 6, Dec. 2016, pp. 1027–45. PubMed Central, https://doi.org/10.1007/s11524-016-0085-2.
[7] Egede, Leonard E., et al. "Modern Day Consequences of Historic Redlining: Finding a Path Forward." Journal of General Internal Medicine, vol. 38, no. 6, May 2023, pp. 1534–37. PubMed Central, https://doi.org/10.1007/s11606-023-08051-4.
[8]Lee, Ceshine. "Feature Importance Measures for Tree Models — Part I." Veritable, 8 Sept. 2020, https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3.

income was also removed in this set of social identifiers. Again, an unhoused individual's urban status serves as a strong predictor of their access to mental health care; mental health care is less specialized and accessible in rural areas.[9] However, because the unhoused population in cities is 38,000, more than ten times the rural homeless population of 3,300, the model's correlation should not be used to delegate resources in a way that does not reflect the population's needs. Another interesting metric is the Asian, Black and Hispanic correlations, which could point to either systemic issues in which communities receive care, or social stigmas around mental health. This is an area where further research is needed.

Predictors	Feature Importance (MDI)
Urban	0.109753
Asian/Pacific Islander	0.085051
Black	0.061704
Age 60+	0.055487
Hispanic	0.054070

5. STABILITY TESTING

Stability testing is necessary to see if results are consistent regardless of what data is being used. By verifying that the model's predictions are reliable, its robustness is ensured for predictions and ultimately policy conclusions.

5.1 Data Perturbation

A random sample of 50% of the final dataframe was chosen to rerun the previous random forest regression model.

Primary Care Shortage

Predictors	Feature Importance (MDI)
Urban	0.093063
Private Insurance	0.90032
Hispanic	0.072734
White	0.066713
Black	0.066135

[9]Morales, Dawn A., et al. "A Call to Action to Address Rural Mental Health Disparities." Journal of Clinical and Translational Science, vol. 4, no. 5, pp. 463–67. PubMed Central, https://doi.org/10.1017/cts.2020.42. Accessed 14 June 2024.

Mental Care Shortage

Predictors	Feature Importance (MDI)
Urban	0.092378
Asian/Pacific Islander	0.080845
Black	0.056617
Female	0.056579
Hispanic	0.055362

5.2 Comparison to Original Model

As seen in the above two tables, the differences in feature importance from the original models are slight. For primary care, private insurance coverage moved to the second most important spot, but the top five identifiers are still the same as the original run of the model. As for mental health care shortage, an interesting change occurred as unhoused female patients replaced the spot held by age 60+ in the original model; however, the change in the actual importance level is less than 0.001, making it marginal. Finally, the accuracy rating of each model increased by 1-2% (78.3% → 79.3% in the primary healthcare model and 72.9% → 74.8% in the mental healthcare model), showing that the model is robust and stable since there was very little variation in the accuracy scores and composition.

6. CONCLUSION

Based on the results of this study, the areas and social identifiers that California's upcoming state policies should focus on are: Hispanic and Black homeless hospitalizations on Medi-Cal, helping people from ages 19-39 get insured, setting acceptable benchmarks for Employer-Sponsored Insurance, and maintaining an adequate number of beds in areas of primary and mental health care shortage.

6.1 Prop 1

During California's 2024 primary election, Proposition 1, titled Bonds for Mental Health Treatment Facilities, was narrowly passed.[10] The bill "approves a $6.4 billion bond to build (1) more places for mental health care and drug or alcohol treatment and (2) more housing for people with mental health, drug, or alcohol challenges."[11]

In the context of this study, California has a current availability of 5,500 beds, and needs 8,000 more units in order to effectively treat mental health issues for unhoused people. Since Prop 1 seeks to build 11,150 new beds using taxpayer money, it would hopefully help address the issue of government facilities that are under-resourced due to lack of beds, as was found in section 3.1.3.

Prop 1 would work by requiring each county to allocate an equal percentage of its tax funds towards mental health for unhoused people. While positive outcomes for unhoused people would certainly happen, it discounts areas of mental and primary care shortage (discussed in sections 4.2 and 4.3) that might lead to wildly inefficient spending disparities in urban and rural areas.

[10] "California Voters Pass Proposition Requiring Counties to Spend on Programs to Tackle Homelessness." PBS NewsHour, 20 Mar. 2024, https://www.pbs.org/newshour/politics/california-voters-pass-proposition-requiring-counties-to-spend-on-programs-to-tackle-homelessness.

[11] Proposition 1 [Ballot]. https://lao.ca.gov/BallotAnalysis/Proposition?number=1&year=2024. Accessed 14 June 2024.