Final Report

AUTHOR
Sean Kim

Introduction:

Access to high quality healthcare is an integral aspect of public health. Ensuring equitable access to such care has been a mission of many public health policies over the years. One measure of access to healthcare is preventable hospitalizations. Individuals with chronic conditions such as diabetes, heart disease, asthma, and chronic obstructive pulmonary disease (COPD) depend on outpatient primary care to keep their conditions under control. With poor access to primary care, these individuals may be forced to travel far for their necessary care or simply deal with their condition until they experience an exacerbation and require hospitalization. These hospitalizations are considered preventable with proper access to appropriate outpatient care. The California Health and Human Services department records data tracking preventable hospitalization rates for ten of the most common conditions managed in the outpatient setting. In addition to preventable hospitalization, CalHHS has also collected data on primary care shortage metrics on a county-, city-, and census area- basis. With these available data, I formulated the following question: Is a shortage of primary care providers in an area associated with increased preventable hospitalizations? I hypothesize that an increased population : primary care provider ratio, indicative of increasing primary care provider shortage will be associated with a higher rate of preventable hospitalizations.

Methods:

Two data sets were selected from the California Health and Human Services (CalHHS) website: "Rates of Preventable hospitalizations for selected conditions" and "Primary care shortage areas". These data sets are open access and downloaded in .csv or .xlsx format and then imported to RStudio for processing. Both datasets were inspected using the "summary", "head", "tail", "str" and "dim" functions for missing or implausible values. The data sets were then combined based on the variable "county". Amador and Calaveras county had observations for "Asthma in younger adults" in which the observations had been replaced by "NA", as the cell size was too small. This is common practice in many studies where such small number of observed cases (less than 5 for example) are removed to protect anonymity of subjects or patients. Considering this, "NA" cells were imputed with 1 arbitrarily. The packages Tidyverse and Dplyr were used for data cleaning. Within each dataset, key variables were identified for analysis.

In the Preventable hospitalizations dataset, PQIDescription, count_ICD10, Population_ICD10, RiskAdjRate_ICD10 were identified. PQI is short for Prevention Quality Indicators, used by the Agency for Healthcare Research and Quality to describe conditions or hospitalizaitons that could be prevented with high quality outpatient care. ICD10 codes represent specific medical diagnoses. Risk-Adjusted-Rates were calculated in this dataset by adjusting the rates accounting for differences in sex, age, and socioeconomic status of the different counties. The dataset also included composite measurements that combined observations for diabetes, acute preventable hospitalizations, chronic preventable hospitalizations, and overall preventable hospitalizations. Count_ICD10 and Population_ICD10 were classified as "character" and needed to be converted to "integer" for analysis. County totals were calculated by taking the mean of observations by ICD 10 code.

For the Primary care shortages dataset, the variables total population, total primary care provider count, population:provider ratio and the ratio score were selected and summarized by county. One county, Alpine, had 0 reported primary care providers. The population:provider ratio was then reported as "NA". This was replaced with 19,000 to match the maximum observed population:provider ratio in the dataset. Primary care provider shortage scores were categorized on a scale of 0-5, with 0 representing minimal shortage of primary care providers and 5 representing the greatest shortage of primary care providers. Specifically, the population:provider ratio for 0 is 0-1000; population:provider ratio for 1 is 1000-1500; 2 is 1500-2000; 3 is 2000-2500; 4 is 2500-3000; 5 is >3000. This Primary care provider shortage score was identified as a key variable for analysis, given the scale (1:5 was more manageable than 700 to 19,000 for population:provider ratios).

In the Exploratory Data analysis "ggplot2" and "kable" were used for data exploration and visualization. Interactive plots were made using the package "Plotly". Key variables were compared with known population data of each county. It was noted that the preventable hospitalization set included a "statewide" observation, which was removed from analysis. In order to visualize the distribution of key variables, box plots and histograms were made of both sets.

0 is 0-1000; 1 is 1000-1500; 2 is 1500-2000; 3 is 2000-2500; 4 is 2500-3000; 5 is >3000

Within the preventable hospitalizations set, a boxplot of risk-adjusted rates (@fig-rates-by-diagnosis) was made for each diagnosis (named "PQI Description" in the original dataset). Risk adjusted rates were used because these rates were determined after adjusting for age, sex, and socioeconomic status of counties, giving a better picture of the healthcare system to be compared between counties. These plots indicate a fairly normal distribution, with a few outliers generally at the top end of the rates, representing counties that had particularly high rates of preventable hospitalizations in those diagnoses.
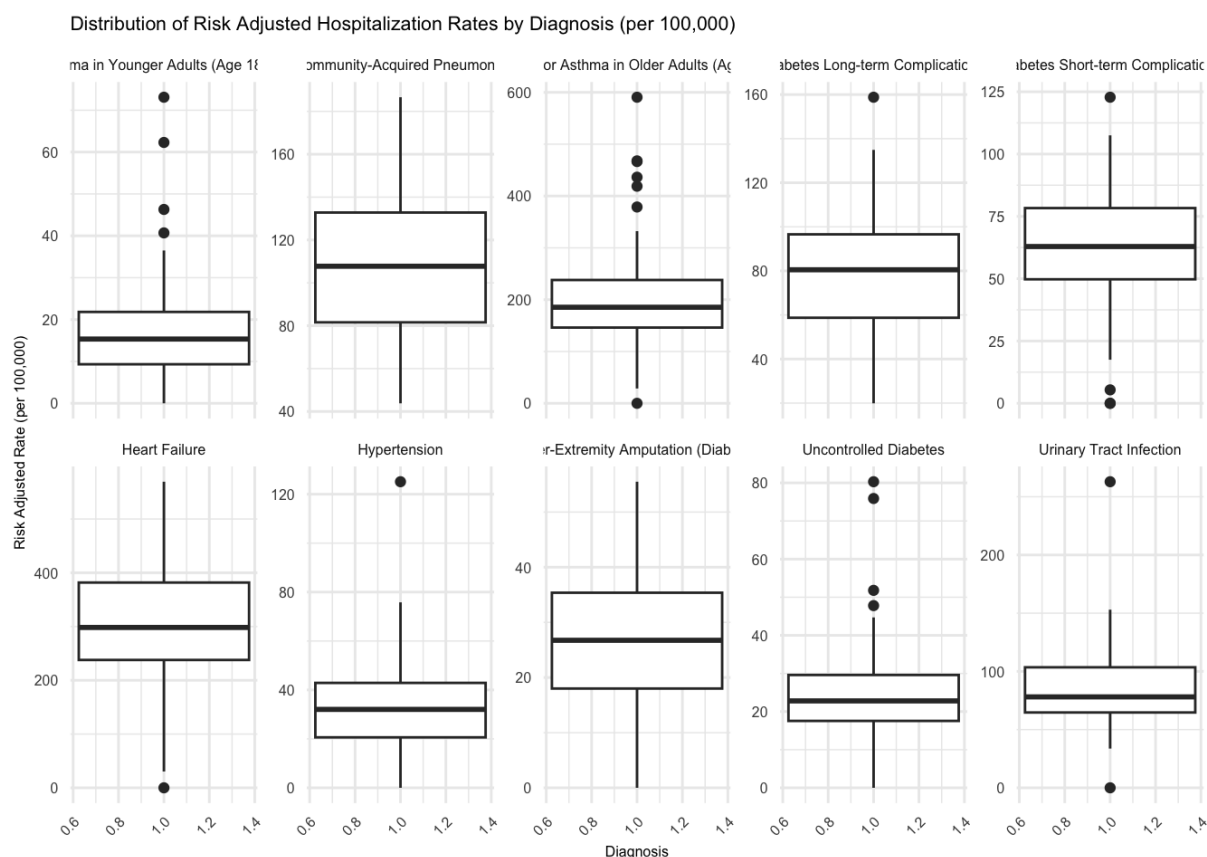

Figure 1: Boxplots showing the distribution of preventable hospitalization rates adjusted for sex, age, and socioeconomic status and displayed in separate facets by diagnosis.

Under the primary care provider shortage set, a histogram was made to illustrate the distribution of provider scores among the different counties – of which there seems to be a mode of 3, a slight left skew, and close to normal distribution (@fig-hist-providerScore). County scores were determined by taking the average of the observed provider scores from different census tracts within each county. Higher scores (4 and 5) indicate a more severe shortage of primary care providers. In one county, Alpine, there are no primary care providers. The boxplot for Population:Provider ratios in California counties also had outliers present (@fig-BoxplotCAProvider), necessitating the use of medians to characterize the ratio most representative of California – which is about 2100 people per primary care provider.
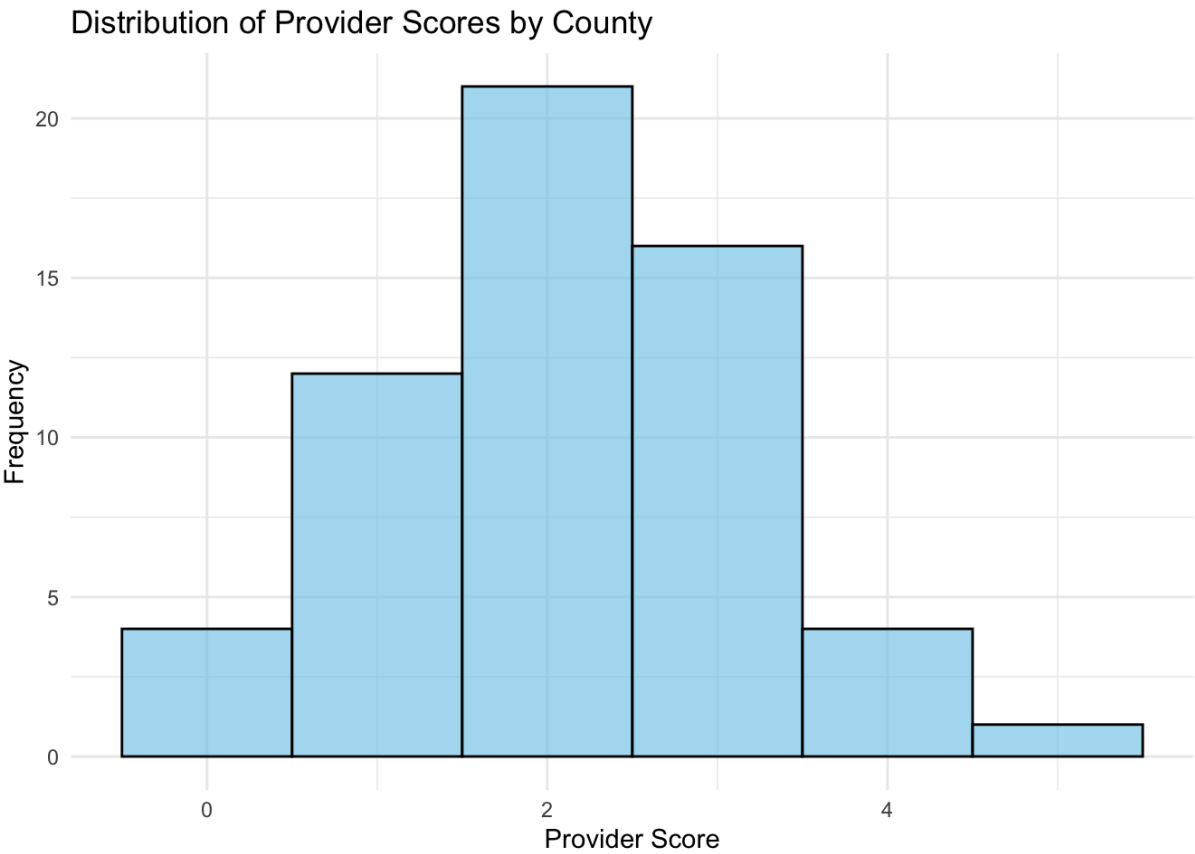
Distribution of Provider Scores by County

Figure 2: distribution of scores of providers by county. 0 is 0-1000 persons per provider; 1 is 1000-1500; 2 is 1500-2000; 3 is 2000-2500; 4 is 2500-3000; 5 is >3000.

?(caption)

Quartiles of Average
Provider Ratio in California

Quartile	Value
0%	Min	738.525
25%	Q1	1377.967
50%	Median	2099.800
75%	Q3	3009.258
100%	Max	18134.225

Quartiles of Provider Ratio in California

```
Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for palette Set2 is 8
Returning the palette you asked for with that many colors
```

```
Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for palette Set2 is 8
Returning the palette you asked for with that many colors
```
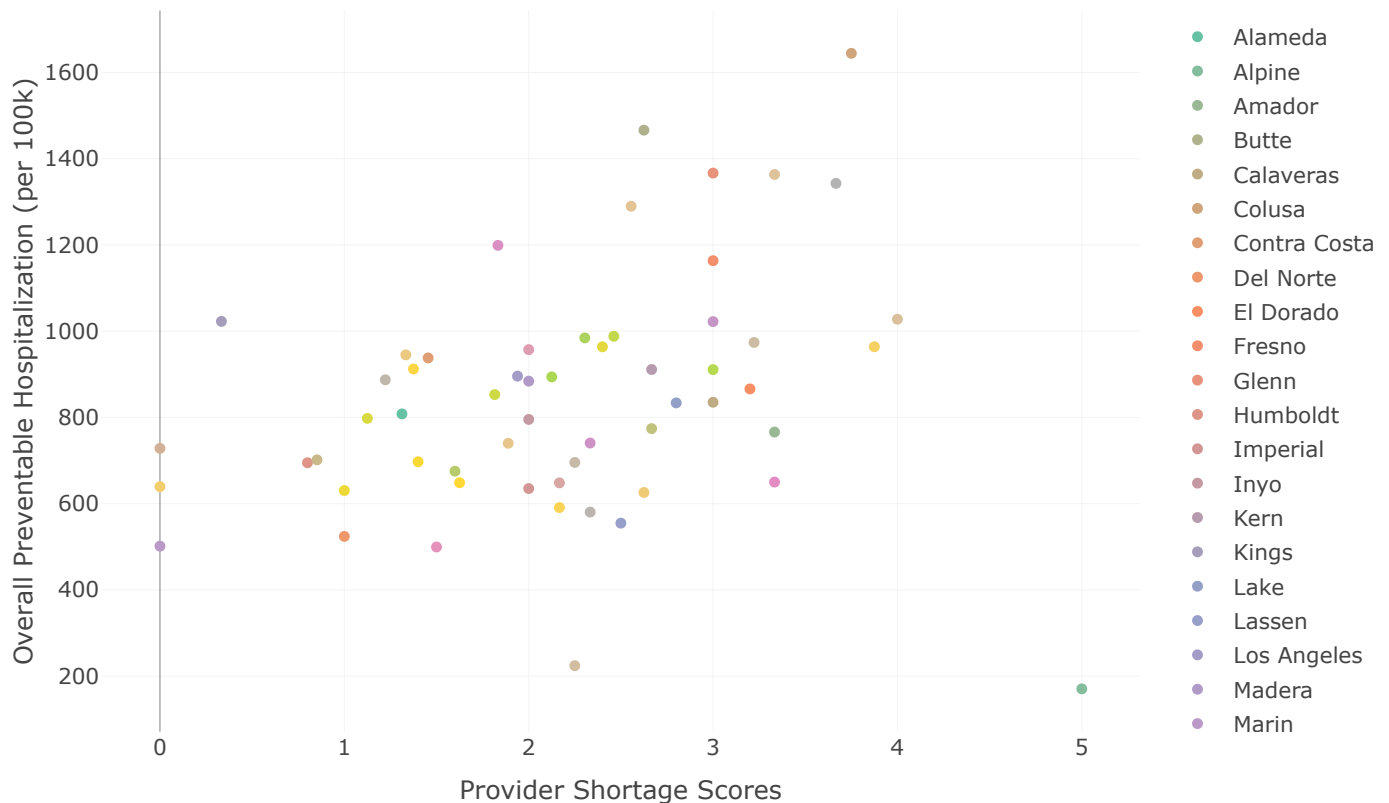

Figure 3: Risk adjusted rate per 100k of overall preventable hospitalizations (adjusted for age, sex, and socioeconomic status by county), versus average county scores for provider shortage scores. Interactive plot, comparison available.

Finishing the exploratory data analysis, the interactive scatter plot (@fig-Interactive-Hospitalizations-by-Shortage) demonstrates the distribution of overall preventable hospitalizations by county. There is added context of provider shortage scores, with 5 indicating severe primary care provider shortage (greater than 1 provider: 3000 residents). Upon visual inspection, there appears to be positive correlation between severity of provider shortage and overall preventable hospitalizations. Further, more granular analysis examining individual diagnoses is warranted.

Results:

Risk adjusted rates for preventable hospitalizations were plotted against population – provider ratios and separated by diagnosis to show a positive correlation between increasing primary care provider shortage and increasing rates of preventable hospitalizations for all diagnoses of interest (@fig-rates-by-providerRatio). Some diagnoses showed stronger correlations than others, indicating conditions more heavily affected by low access to primary care providers.

Preventable hospitalization rates were also examined by groups of diagnoses, aggregating diabetes complications, acute reasons for hospitalization, exacerbation of chronic conditions, and overall measurements (@fig-Rate-vs-score). These measures were compared to the provider shortage score, which ranges from 0 – 5, with increasing scores meaning greater shortage of providers. Comparisons were found to be positively correlated on these composite levels as well.

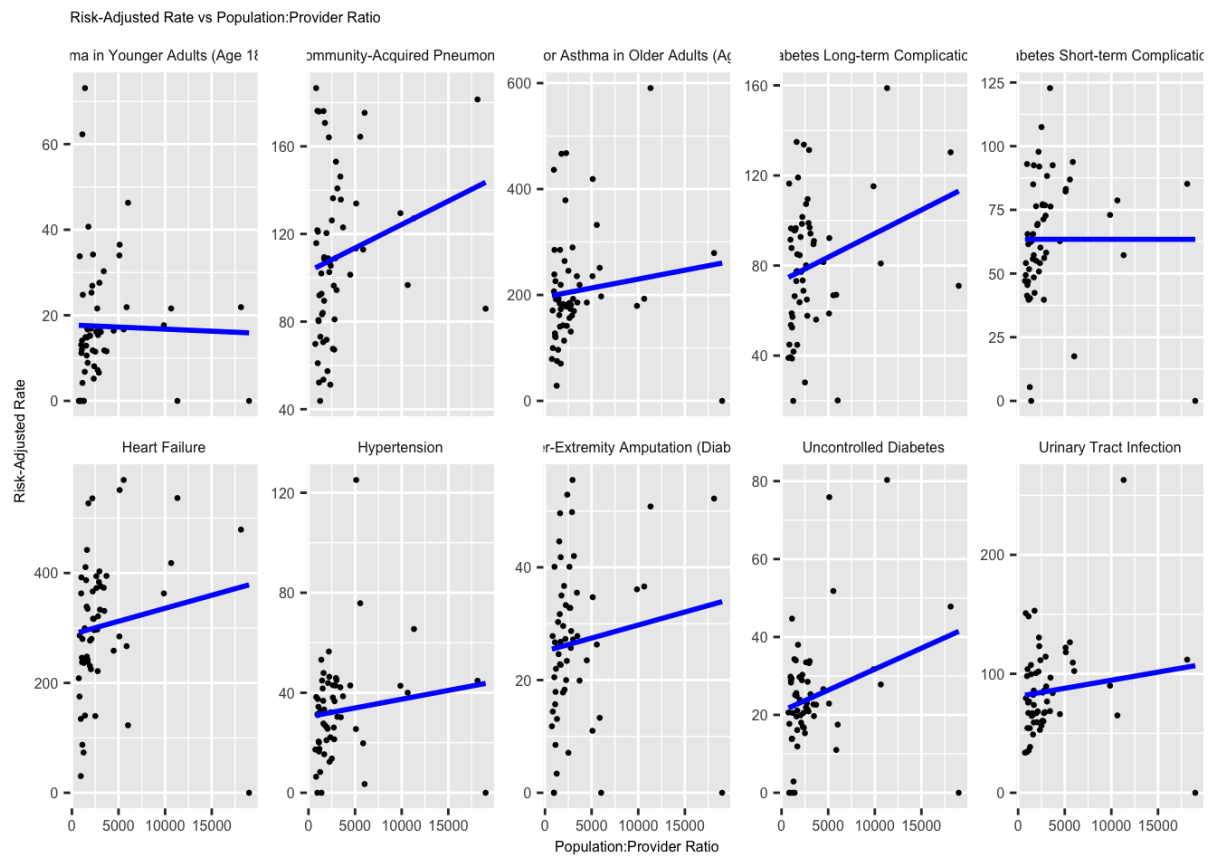`geom_smooth()` using formula = 'y ~ x'

Figure 4: Plots of preventable hospitalization rates vs population:provider ratio separated by diagnosis.

```
`geom_smooth()` using formula = 'y ~ x'
```

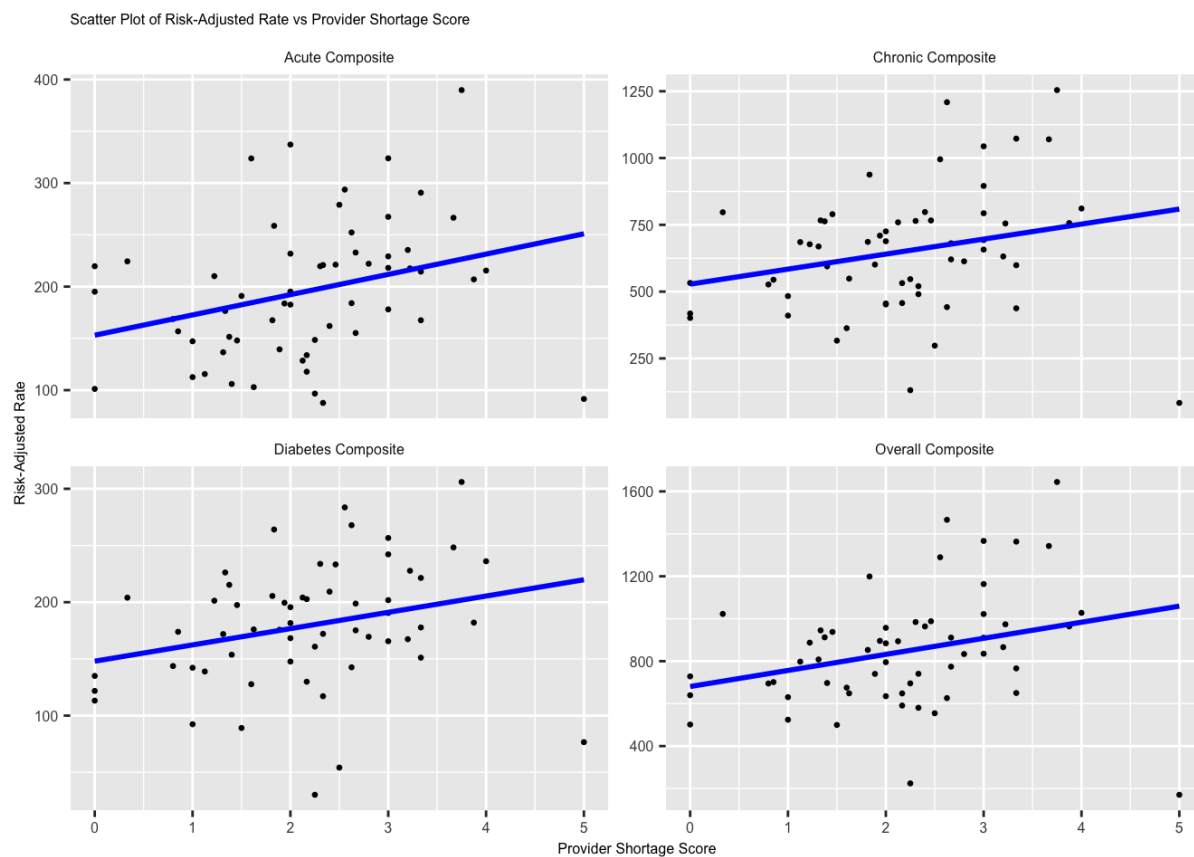Scatter Plot of Risk-Adjusted Rate vs Provider Shortage Score

Figure 5: Rate of preventable hospitalizatoin vs Provider shortage score based on composite measures. Diabetes composite is made up of short term complications, long term complications, lower extremity amputation due to diabetes, and uncontrolled diabetes. Chronic composite is made up of diabetes composite measures, COPD and Asthma in older and younger adults, hypertension, and heart failure admissions rates. Acute composite is made up of community acquired pneumonia and urinary tract infection admissions rates.

Summary Statistics:

Counties Ordered by Highest Composite Risk-Adjusted Rates

County	Adjusted Rate of Preventable Hospitalization	Provider Shortage Score
Colusa	1644.4	3.7500000
Butte	1466.2	2.6250000
Glenn	1366.7	3.0000000
Sutter	1363.4	3.3333333
Yuba	1342.8	3.6666667
Stanislaus	1289.7	2.5555556
Merced	1199.1	1.8333333
Fresno	1163.5	3.0000000
Tehama	1027.8	4.0000000
Kings	1022.7	0.3333333

Top 10 Counties by Provider Score

County	Adjusted Rate of Preventable Hospitalization	Provider Shortage Score
Alpine	170.5	5.000000
Tehama	1027.8	4.000000
Shasta	963.9	3.875000
Colusa	1644.4	3.750000
Yuba	1342.8	3.666667
Amador	766.1	3.333333
Modoc	650.1	3.333333
Sutter	1363.4	3.333333
Tulare	973.8	3.222222
El Dorado	866.1	3.200000

Conclusions:

The top ten counties were taken in terms of adjusted rates of preventable hospitalization and provider shortage scores. Notable counties that appeared in the top 10 on both lists include Shasta, Colusa, Yuba, and Sutter. These four counties may represent target areas that experience both the highest rates of preventable hospitalization and the greatest primary care provider shortages.

These datasets had some noteworthy limitations. The Primary Care provider shortage dataset only included data from the year 2020, making it difficult to draw conclusions about any time-course. It would be interesting to see how the changes in primary care provider availability associate with changes in preventable hospitalizations. The Preventable hospitalizations dataset was also limited in its scope. More granular analysis of individual diagnoses by county was difficult due to low rates of hospitalization in some counties. For example, there were 0 hospitalizations due to asthma in young adults in several of the smaller counties in 2020.

Preventable hospitalizations indicate areas of improvement in public health, particularly in access to care. These data show a positive association between shortage of primary care providers and preventable hospitalization rates. Greater investment in primary care, particularly in areas of shortage, can potentially improve these rates and citizens' overall health.

Links to Datasets and Data Dictionaries:

Primary Care Shortage Areas:

- https://data.chhs.ca.gov/dataset/primary-care-shortage-areas-in-california/resource/0ba7c904-2302-400a-ba27-b8e8e5c1ab4a

Preventable Hospitalizations:

- https://data.chhs.ca.gov/dataset/rates-of-preventable-hospitalizations-for-selected-medical-conditions-by-county/resource/1f699c45-f52f-408e-a8f2-87e537aea82d?inner_span=True