R Project Milestone 6

Moyra Rasheed, Courtney Coon, Jarett Maycott

2022-11-28

Problem Statement

Rural hospitals are struggling to stay open throughout the country and the state of California ranks #4 in rural hospital closures (citation). When rural hospitals close, there is a subsequent rise in mortality (citation). This is why the California Department of Public Health, Office of Health Equity is excited that a new policy has just been created to fund a public-private partnership for healthcare facility improvement in 5 of California's rural counties that have received minimal funding from the Department of Health Care Access and Information (HCAI) over the past 5 years. Here we have analyzed 5 variables in order to determine which 5 counties are the best targets for the development fund proposals.

Methods

Data Sources

To make recommendations on target counties, we used variables from three open-access, State of California data sources which are described below.

California demographics by County Includes data on population characteristics by county including average population density, counts by race/ethnicity, sex, median age, and housing data.

Years and/or dates of data

Data was collected in 2012.

Description of cleaning and creating new variables

To prepare the data, we:

1. Data cleaning

- make "County" the first (left-most) column in the data set for merging

2. Data filtering

- subset data to only include needed data

3. Data summarizing

- *NEW VARIABLE*: categorize Median Age into "High Priority" (counties with Median Age above the third quantile), "Medium Priority" (counties with Median Age between the first and third quantiles), and "Low Priority" (counties with Median Age below the first quantile)
- *NEW VARIABLE*: categorize Population per Square Mile (hereafter Population Density) into "High Priority" (counties with Population Density above the third quantile), "Medium Priority" (counties with Population Density between the first and third quantiles), and "Low Priority" (counties with Population Density below the first quantile)
- *NEW VARIABLE*: create ratio of renting households as a fraction of total households.
- *NEW VARIABLE*: categorize Percent of Household that were Renters (hereafter Percent Renters) into "High Priority" (counties with Percent Renters above the third quantile), "Medium Priority" (counties with Percent Renters between the first and third quantiles), and "Low Priority" (counties with Percent Renters below the first quantile)

California Department of Public Health Care: Death Profiles by County Includes data on mortality events stratified by age, sex, race, hospital department, and residence status of the deceases relative to the hospital where the mortality event occurred.

Years and/or dates of data

The data set contains data from 2014 through 2020.

Description of cleaning and creating new variables

To prepare the data, we:

1. Data cleaning

- made all column names lowercase
- replaced NAs with zeros, as instructed

2. Data filtering

- filtered data in "strata_name" to only include data about the Total Population (removed age, sex, race and other stratifiers)
- filtered out non-chronic diseases (all cause, assault, accidents, influenza and self-harm)
- removed data from 2014

3. Data summarizing

- *NEW VARIABLE*: sum all chronic disease mortality events by county (both occurrence and residence data)
- *NEW VARIABLE*: [after merging data sets] divide summed chronic disease mortality by average population size (from the demographic data set) to create a rate of chronic disease mortality per person for each county
- *NEW VARIABLE*: categorize Chronic Disease Mortality Rates (hereafter Mortality) into "High Priority" (counties with Mortality above the third quantile), "Medium Priority" (counties with Mortality between the first and third quantiles), and "Low Priority" (counties with Mortality below the first quantile)

Department of Health Care Access and Information: Total Construction Cost of Healthcare Projects Includes number of projects and total spending on hospital projects funded by Department of Health Care Access and Information (HCAI) by county, date, and project status ("in review", "pending construction", "in construction" and "in closure.")

Years and/or dates of data

The original data set contains data from 10-14-2013 through 08-11-2022.

Description of cleaning and creating new variables

To prepare the data, we:

1. Data cleaning

- made all column names lowercase
- replaced spaces with underscores
- removed the numbers in front of the County names
- removed dollar signs from Total Costs
- changed date format

2. Data filtering

- removed projects listed as "in review" because we wanted to prioritize locations with the fewest projects that were at any stage of completion
- removed data from 2013 and 2014 to match other available data

3. Data summarizing

- average Total Costs in each category for each county to reduce over-counting
- *NEW VARIABLE*: sum the Total Costs over the 4 categories ('in closure', 'in construction', or 'pending construction') for each county
- *NEW VARIABLE*: categorize Total Costs into "High Priority" (counties with Total Costs below the first quantile), "Medium Priority" (counties with Total Costs between the first and third quantiles), and "Low Priority" (counties with Total Costs above the third quantile)

Analytic methods

We decided to approach the ranking of counties with two techniques to examine the robustness of our final recommendations. Our first approach was to use two scatterplots that build off each other. Plot 1 has just the 3 demographic variables of interest: fraction of renters, median age of the residents, and population density. For Plot 2, we collapsed the demographic data into a single *NEW VARIABLE* by coding giving 2 points for values identified as "high priority," 1 point for values identified as "medium priority" and 0 points for "low priority" values for each county. For example, Amador County is a high priority in regards to the median age and population density of its residents (2 points for each variable), but a low priority based on fraction of renters (0 points), for a total demographic ranking score of 4 points. This new variable was used to color code the dots in Plot 2 where we graph Previous Funding and Chronic Disease Mortality Rate. Plot 2 highlights the counties that we would consider recommending for future funding because it includes all 5 variables of interest.

Our second approach was to use a table with a *NEW VARIABLE* to rank the counties on all 5 variables of interest simultaneously. The ranking variable uses a similar method as above except that the point values were not universal. We decided that the Chronic Disease Mortality Rate, Population Density and Median Age were more valuable than the Renter Ratios and Previous Investments because (1) Renter Ratio is inversely correlated with Median Age (the older a person gets, the more likely they are to own a home) and (2) because we believe it is more equitable to fund hospitals that appear to be in need of funding regardless of prior funding. For the 3 "more valuable" variables, we gave 4 points for "high priority" values, 2 point "medium priority" values, and 0 points for "low priority" values. For the 2 less important variables, we gave 2 points to values identified as "high priority," 1 point for values identified as "medium priority" and 0 points

for "low priority" values for each county. For example, Siskiyou County is a high priority based on median age (4 points), population density (4 more points) and chronic disease mortality rates (4 points). It was a medium priority based on total previous investment (1 point) and a low priority based on fraction of renters (0 points) for a total ranking score of 13 points. Counties were then ranked on their total ranking score.

Results

Figure 1: **High priority counties based on demographic data only**

Priority is given to counties with: (1) renter ratios above the median of all counties (median = 39%), (2) median population age above the median for all counties (median = 37 years old), and (3) population densities below the third quantile for all counties (High priority is below the 1st quantile = 26 people per square mile; Low priority is above the 3rd quantile = 333 people per square mile). Each dot represents a county in California. Counties that are named on the figure are those meeting all of the high priority demographic criteria.

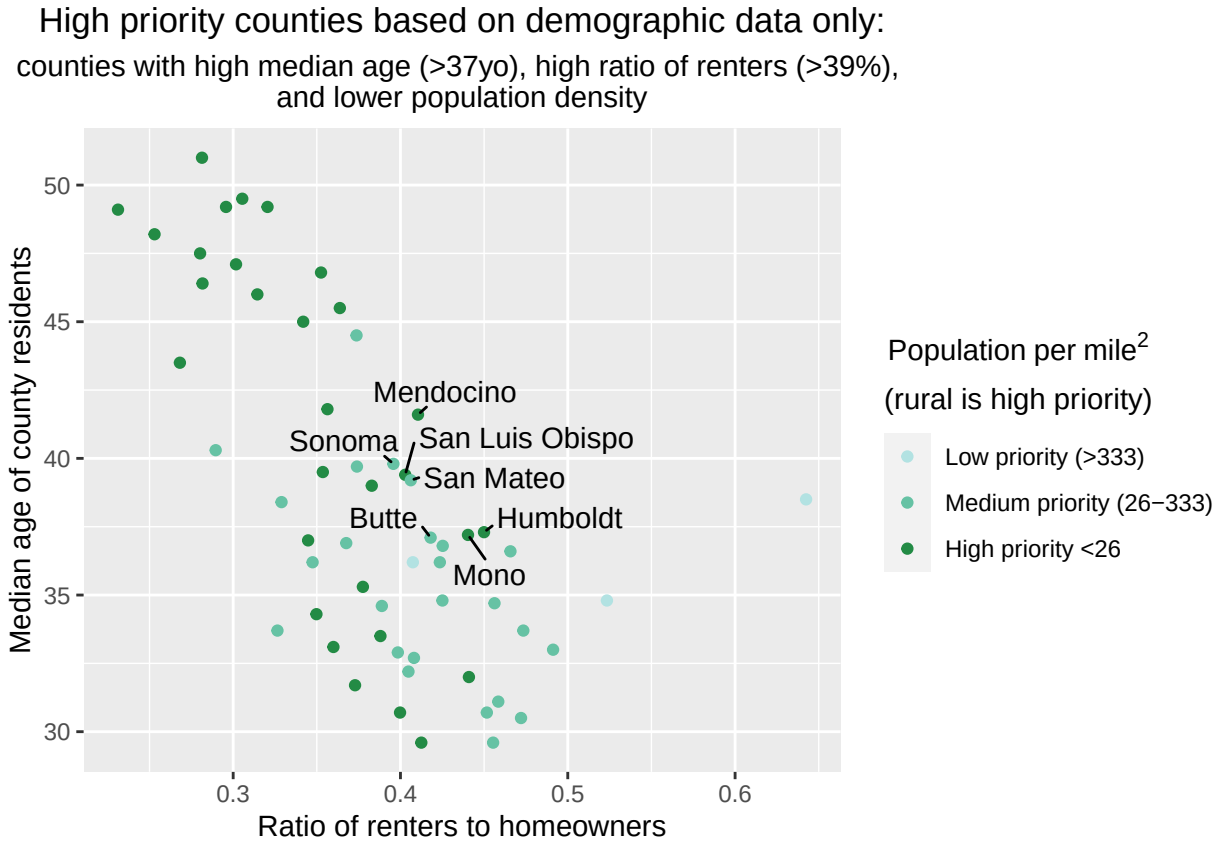
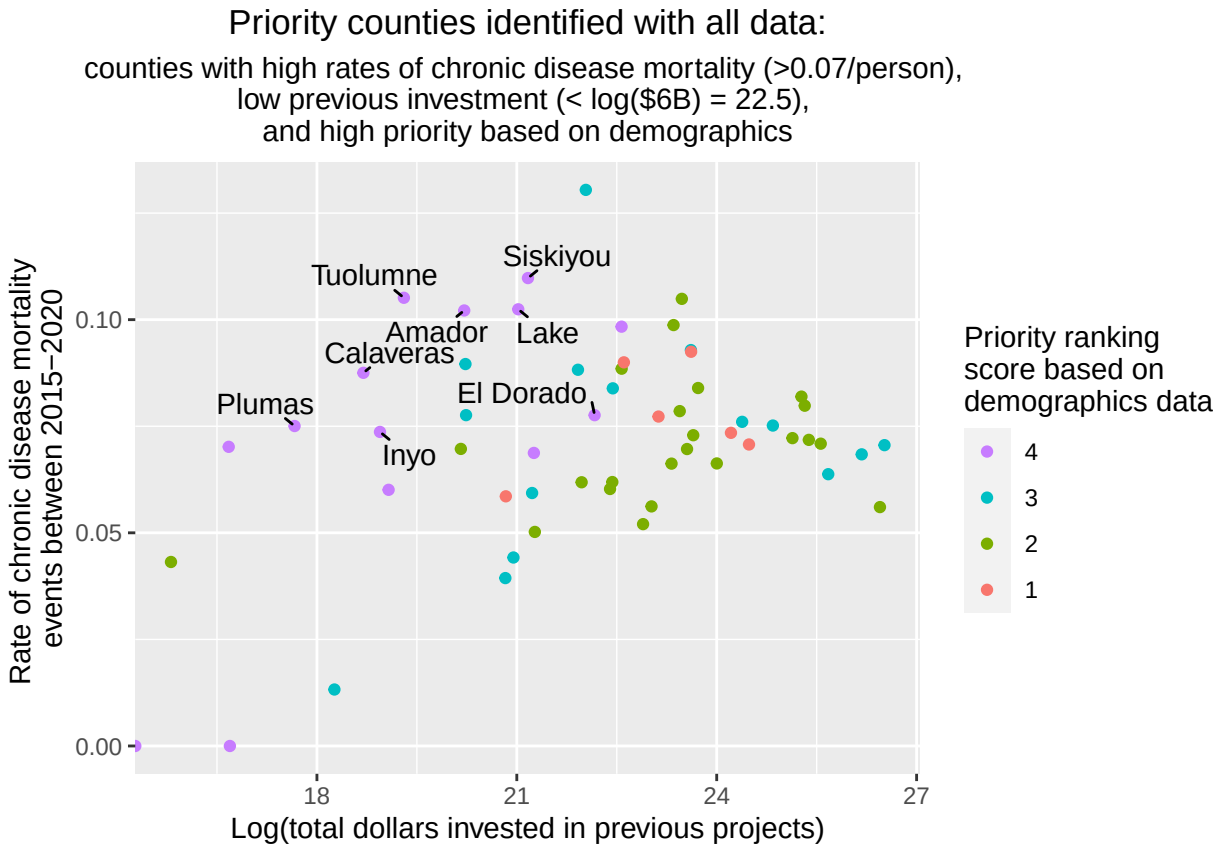


Figure 2: **High priority counties using all variables of interest.**

Color coding is based on the demographic data presented in Plot 1. A County with higher priority demographic values received a higher score (score of 4 has the highest priority). Also in the Plot is the total dollars previously invested in prior projects with higher priority for counties below the median previous investment (log($5,961,782,208) = 22.50864) on the x-axis, and, on the y-axis, rate of chronic disease mortality events between 2015-2020, with higher priority counties having higher than the median rates of chronic disease mortality (median = 0.07 cases of mortality per person).



To chose the top 5 Counties out of the 8 the graph identifies is subjective and depends on whether one wants to prioritize rate of chronic disease mortality or previous spending. As mentioned in the Analytic Methods section, our group believes rate of chronic disease mortality is more valuable so, using this graphical method, we would recommend: Tuolumne, Amador, Calaveras, Siskiyou, and Lake Counties for future funding.(If previous spending was more valuable, Tuolumne, Amador, Calaveras, Plumas, and Inyo would be prioritized).

Table 1: Top 10 Counties ranked by need for oshpd projects.

County	Previous spending on projects		Population density		Median age of population		% population that are renters		Chronic disease mortality burden	
Amador	0.1021536	High priority	63.288340	High priority	48.2	High priority	0.2530030	High priority	598970736	Low priority
Calaveras	0.0875314	High priority	44.582939	High priority	49.1	High priority	0.2311765	High priority	131848234	Low priority
Tuolumne	0.1051309	High priority	24.304973	High priority	47.1	High priority	0.3017241	High priority	242129946	Low priority
Lake	0.1024321	High priority	49.082334	High priority	45.0	High priority	0.3418713	Medium priority	1347450993	Low priority
Nevada	0.0983582	High priority	102.564339	High priority	47.5	High priority	0.2802273	Medium priority	6352716267	Low priority
Siskiyou	0.1097566	High priority	7.120891	High priority	46.8	High priority	0.3525250	Medium priority	1558949981	Low priority
Inyo	0.0736661	High priority	1.819773	High priority	45.5	Medium priority	0.3637719	High priority	169160700	Low priority
Mariposa	0.0701707	High priority	12.613887	High priority	49.2	Medium priority	0.3205512	High priority	17474756	Low priority
Plumas	0.0750500	High priority	7.653217	High priority	49.5	Medium priority	0.3054473	High priority	46955168	Low priority
Tehama	0.0895902	High priority	21.523312	Medium priority	39.5	High priority	0.3535995	High priority	610226591	Low priority
El Dorado	0.0775971	High priority	102.156840	High priority	43.5	Medium priority	0.2681742	Medium priority	4239088028	Low priority
Humboldt	0.0928689	High priority	38.062105	Medium priority	37.3	High priority	0.4499474	Medium priority	17981394511	Low priority
Modoc	0.0687366	High priority	2.329272	High priority	46.0	Medium priority	0.3144685	Medium priority	1703663257	Low priority
San Luis Obispo	0.0882523	High priority	81.815416	Medium priority	39.4	High priority	0.4028388	Medium priority	3302001201	Low priority
Shasta	0.1304416	High priority	46.480517	Medium priority	41.8	High priority	0.3563671	Medium priority	3715362195	Low priority

Table:

JARETT: add caption

```
## # A tibble: 15 x 12
##    county      relative_chroni~ pop12_sqmi med_age renter_ratio summed_total_co~
##    <chr>                  <dbl>      <dbl>   <dbl>        <dbl>            <dbl>
##  1 Amador                 0.102       63.3     48.2        0.253       598970736.
##  2 Calaveras              0.0875      44.6     49.1        0.231       131848234.
##  3 Tuolumne               0.105       24.3     47.1        0.302       242129946
##  4 Lake                   0.102       49.1     45          0.342      1347450993.
##  5 Nevada                 0.0984     103.      47.5        0.280      6352716267.
##  6 Siskiyou               0.110        7.12    46.8        0.353      1558949981.
##  7 Inyo                   0.0737       1.82    45.5        0.364       169160700.
##  8 Mariposa               0.0702      12.6     49.2        0.321        17474756
##  9 Plumas                 0.0750       7.65    49.5        0.305        46955168
## 10 Tehama                 0.0896      21.5     39.5        0.354       610226591.
## 11 El Dorado              0.0776     102.      43.5        0.268      4239088028.
## 12 Humboldt               0.0929      38.1     37.3        0.450     17981394511.
## 13 Modoc                  0.0687       2.33    46          0.314      1703663257
## 14 San Luis O~            0.0883      81.8     39.4        0.403      3302001201.
## 15 Shasta                 0.130       46.5     41.8        0.356      3715362195.
## # ... with 6 more variables: relative_chronic_dis_mort_CAT <fct>,
## #   pop12_sqmi_CAT <fct>, med_age_CAT <fct>, summed_total_cost_CAT <fct>,
## #   renter_ratio_CAT <chr>, rank <dbl>
```

The ranking variable described in the Analytic Methods identifies 3 Counties as top priorities - Amador, Calaveras, and Tuolumne - and 3 more Counties as second-rank priorities - Lake, Nevada, and Siskiyou.

Discussion Based upon our inital interpretation of data ,we were able to differentiate 8 counties that would benefit from greater healthcare funding: Amador,Calveras, Tuolumne,Siskiyou,,Plumus, Inyo,El Dorado and Lake Counties. We initially established 5 criteria during our analysis including chronic disease mortality, median age, population density, total cost projects and renter's ratio. Based upon our initial plot, we narrowed the above counties to 5 ,placing greater weight on mortality, age and population density.We placed less significance on renters ratio and project costs, with the assumption that higher median age are more likely to be owners rather than renters and cost allocation is more likely to be directed to those counties with greater healthcare needs. Based upon the above , we narrowed our county allocation to the following: Tuolumne, Amador, Calaveras, Siskiyou,and Lake Counties. (5 counties). Both out plot and table are reflecting relatively similar data, other than 1 variation.In our analytical data table, Nevada county ranks

higher than Siskiyo and Inyo(but this is not reflected in our plot)because although Nevada scored higher points,the higher score was established from higher renters ratio and project costs, rather than the 3 criteria we laid emphasis on.