Final Project

Jeremias Castillo

2023-12-13

Introduction:

Do urban areas having a higher total percent of high-risk individuals for COVID-19 translate to larger number high-risk individual per ICU bed across the United States? This study is conducted to better investigate the main hypothesis that, on average, urban areas in the United States with higher rates of individuals at high risk for COVID-19 means more high-risk individuals per ICU bed in hospitals. Urban areas have historically been densely populated with disadvantaged minorities predisposed to worsening health disparities not being abridged.

Exacerbated health disparities can be better understood through the renowned American epidemiologist and Harvard T.H. Chan School of Public Health, Professor Nancy Krieger's (1994) theoretical framework of ecosocial theory of disease distribution. It claims that health outcomes are driven by numerous other factors known as the social determinants of health and not by the assumption of inherent ailment that erroneously follows minority demographics in the health sector. On the basis of this theory, the hypothesis under investigation is sound in assuming that there might be some sort of relationship between higher rates of people at high risk for COVID-19 and the number of high-risk individuals per ICU bed in hospitals across the urban U.S. as there might be fewer resources available in said communities to adequately address the novel pandemic.

Data Section:

The data set used to perform this cross-sectional study is called "covid-geography", saved in `mmsa-icu-beds.csv`, which merges data from two other data sets, being the Behavior Risk Factor Surveillance System by the CDC, surveying the status of American health, and another by the Kaiser Family Foundation, compiling available urban hospital reports (which could be considered somewhat of a census attempt). The independent variable of interest in this study is the total percent of high-risk individual for COVID-19 in U.S. urban centers. The dependent variable of interest is the number of high-risk individuals for COVID-19 per ICU bed in urban U.S. hospitals.

```r
library(tidyverse)
```

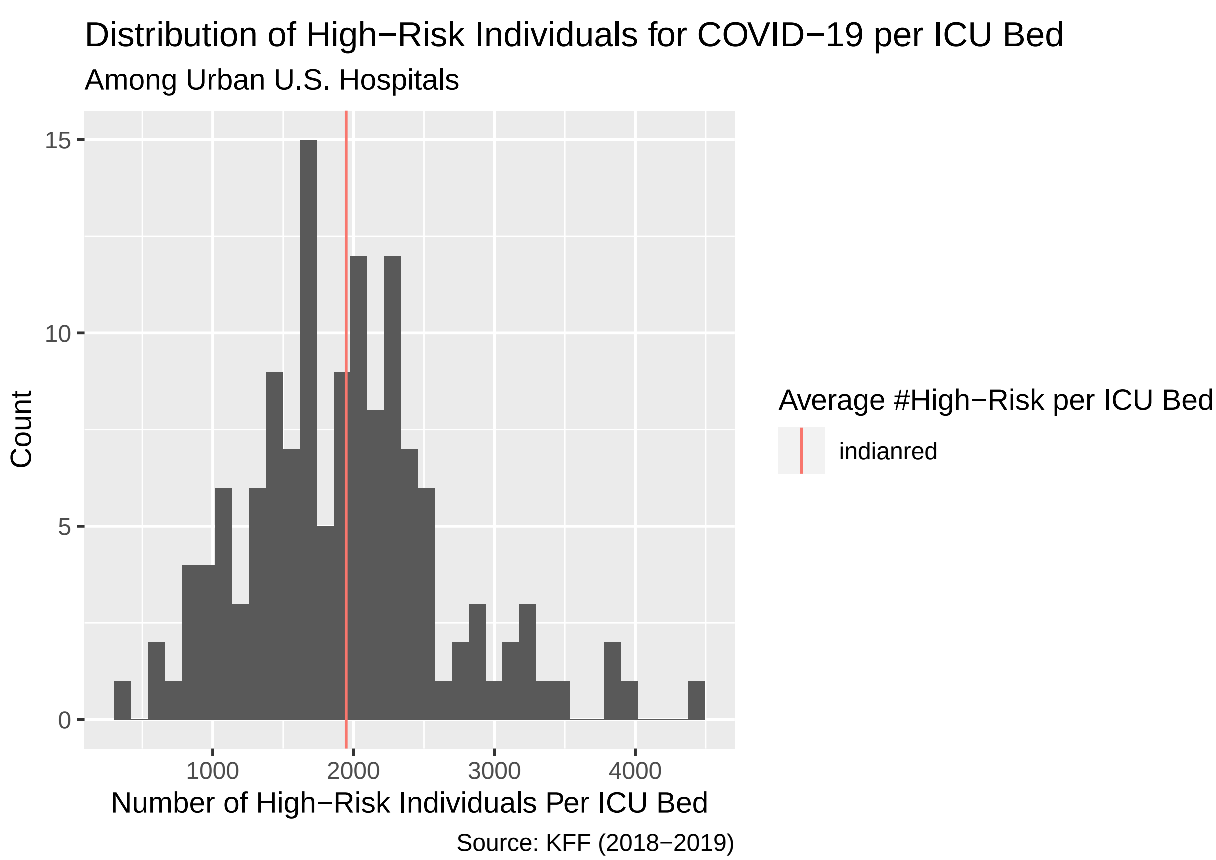
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
covid <- read_csv("data/mmsa-icu-beds.csv")
```

```
## Rows: 136 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): MMSA, total_percent_at_risk
## dbl (5): high_risk_per_ICU_bed, high_risk_per_hospital, icu_beds, hospitals,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
covid |>
  ggplot(mapping = aes(x = high_risk_per_ICU_bed)) +
    geom_histogram(bins = 35) +
    geom_vline(mapping = aes(xintercept = mean(high_risk_per_ICU_bed, na.rm = TRUE), color = "indianred
    labs(
      title = "Distribution of High-Risk Individuals for COVID-19 per ICU Bed",
      x = "Number of High-Risk Individuals Per ICU Bed",
      y = "Count",
      subtitle = "Among Urban U.S. Hospitals",
      caption = "Source: KFF (2018-2019)",
      colour = "Average #High-Risk per ICU Bed"
        )
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```



Shown above is the distribution of the number of high-risk individuals for COVID-19 per ICU bed among urban U.S. hospitals. The average number of people at high risk for COVID-19 per ICU bed is approximately 1947. This study is generalizable to populations within a similar range, on average, of high-risk individuals per ICU bed as this is where most data used is concentrated and best put to use without extrapolation.

Results Section:

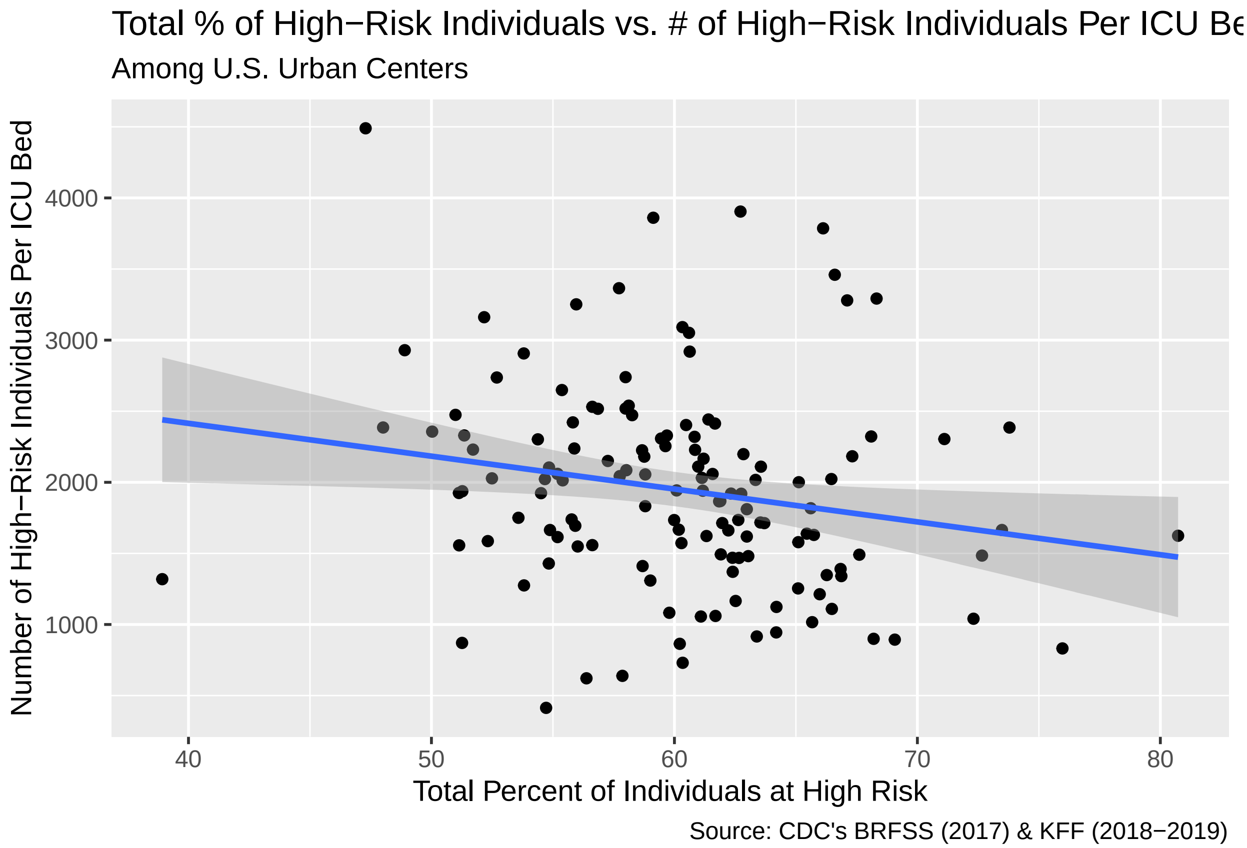
```
covid$total_percent_at_risk <- gsub("%", "", paste(covid$total_percent_at_risk))

covid |>
  ggplot(mapping = aes(x = as.numeric(total_percent_at_risk), y = high_risk_per_ICU_bed)) +
    geom_point() +
    geom_smooth(method = "lm", se = TRUE) +
    labs(
      title = "Total % of High-Risk Individuals vs. # of High-Risk Individuals Per ICU Bed",
      x = "Total Percent of Individuals at High Risk",
      y = "Number of High-Risk Individuals Per ICU Bed",
      subtitle = "Among U.S. Urban Centers",
      caption = "Source: CDC's BRFSS (2017) & KFF (2018-2019)"
        )
```

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

## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



The scatter plot above includes a graphed least-squares regression line that depicts how as the total percent of people at high risk for COVID-19 increase, the number of high-risk individuals per ICU bed decreases in urban centers across the U.S. This relationship might be an indicator that urban centers with greater total rates of high-risk individuals in their populations have a lot more resources allocated to them that allow them to minimize counts for individuals per ICU bed in their hospitals. They might be allotted more beds in their ICUs, meaning better capacity and ability to handle the pandemic. On the other hand, urban centers with lower rates of individuals at high risk in their populations might have a much more fragile system due to lower capacity. This finding highlight a possible disparitiy in resources available to healthcare facilities in urban center with a low rates of individuals at high-risk, especially during pandemic, epidemics, or outbreaks similar to COVID-19. These urban center were expected, under the hypothesis investigated, to have a much lower number of high-risk individuals, yet the number is higher. As a direct association between this assumption and viewed correlation cannot be immediately made, it encourages further research of the subject.

```
covid$total_percent_at_risk = as.numeric(as.character(covid$total_percent_at_risk))

fit_covid <- lm(high_risk_per_ICU_bed ~ total_percent_at_risk, data = covid)

var_labels <- c(
  "(Intercept)" = "Y-Intercept",
  "total_percent_at_risk" = "Total Percent At High-Risk"
              )

modelsummary::modelsummary(list(Regression = fit_covid),
                           statistic = c("s.e. = {std.error}",
                                         "p = {p.value}"),
```

	Regression
Y-Intercept	3338.832
	s.e. = 605.263
	p = <0.001
Total Percent At High-Risk	−23.102
	s.e. = 9.998
	p = 0.022
Num.Obs.	135
R2	0.039
R2 Adj.	0.031

```
           gof_map = c("nobs", "r.squared", "adj.r.squared"),
           coef_map = var_labels)
```

As seen in the linear regression above, the y-intercept of the data set available is approximately 3,338.832 high-risk individuals per ICU bed available. Meaning that when the percentage individuals at high risk is zero, this would be the value of individuals at high risk from COVID-19 per ICU bed available. Of course, this is an extrapolation from what is available in the data set being worked with. The slope of the linear regression is approximately -23.102; this means that for every one percent increase in the total percent of individuals at high risk, there is a decrease of 23.102 high-risk individuals per ICU bed available.

Both the left and right extremes on the scatter plot, however, show an increase in the standard error; the extrapolation occurring due to a lack of data is explanatory for the high standard error seen here for the y-intercept of the linear regression done. The extrapolation here is also causing our r-squared to be heavily penalized. Even so, focus should be drawn to the range where standard error is minimized on the linear model. This is what is representative of a statistically significant trend as shown by the slope's p-value of 0.022 with a standard alpha level of 0.05.

Conclusion:

Though undeniable that there are many health disparities among minorities who mostly live in urban centers, it seems that, on average, hospitals have the ICU capacity to serve the communities they are in when they have high rates of high-risk individuals for COVID-19. Unfortunately, circumstances are not better for urban centers with lower rates of high-risk individuals as they have a higher density of people at high risk per ICU bed in their hospitals. These results, however, could be better defined if more data is collected on this group and merged into this data set as it is mostly based on extrapolation which would require more resources and time as a large-scale project taken on by a team. As with most issues in public health, even those addressed by Nancy Krieger, there is a significant lack of data in the sector that impedes more sophisticated studies from being made, but this study also serves as a starting point to continue this investigation. For this reason, the null hypothesis under investigation in this study could be found to have convincing evidence in its favor, but based on the analyses here with all concerns aside, including aforementioned extrapolation, there is statistically significant evidence in its favor, as observed in the linear regression conducted.