HS256F - 1 – Healthcare Data Analytics and Data Mining

Professor Moaven Razavi

January 30, 2022

HOMEWORK 1

ANALYSIS OF THE NATIONAL PLAN AND PROVIDER ENUMERATION SYSTEM (NPPES) DATABASE

GROUP 4

Christianah Adeoya

Cyndi Ng

Siddhartha Kumar

Zilin Luo

Introduction

The Centers for Medicare and Medicaid Services (CMS) developed the National Plan and Provider Enumeration System (NPPES) as a means of assigning unique identifiers to health care providers and health plans in line with the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The general idea was to improve the efficiency and effectiveness of transmitting electronic health information (*NPPES*, n.d.). These unique identifiers assigned by the CMS were called National Provider Identifiers (NPIs) are unique ten-digit numeric identifiers assigned to all HIPAA providers, individuals and organizations alike (Medicare Learning Network, 2021).

This report seeks to explore, if there exists, any gender difference within eight states in the establishment of solo medical practices, as well as in the choice of a physician's specialty, i.e., to see if one gender is more pre-disposed to low risk - low reward specialties or high risk - high reward specialties than the other, and vice versa. The eight states this report focuses on are Hawaii, Michigan, Minnesota, Mississippi, New York, Oklahoma, South Dakota and Tennessee.

A combination of statistical packages were used in this analysis including STATA/SE 17.0 and Python, and subsets of the main dataset were created as necessary. In exploring the relationships this report is anchored on, Fisher's exact test was employed and the results discussed with regard to statistical significance.

Part 1 - Identification Of Healthcare Providers Of Report Contributors

From the most recent version of the dataset, the NPI and state in which the health care provider for each contributor was pulled and collated as follows:

Contributor's Last Name	NPI	State of Healthcare Provider's First License
Adeoya	1477529675	IL
Kumar	1275536401	MI
Luo	1972507325	MA
Ng	1801936497	MA

Table 1: List of state of the first license of report contributors

Part 2 - Gender Difference in Practicing as a Sole Proprietor

This portion of the analysis sought to explore the difference, if any, in the establishment of solo practices in the aforementioned eight states in the United States using the Fishers exact statistical test.

Method: Using STATA/SE 17.0, the data for the eight states in question were filtered and cropped out to create a sub database. The individual entities were also filtered and non-individual entities excluded from the new subset. In order to ensure a 2x2 table for the Fisher's exact test, all observations with values other than 'Y' or 'N' for the sole proprietorship variable were dropped. The Fisher's exact test was then run for the gender variable as against sole proprietorship in a 2x2 table.

Result and Analysis: The result obtained from the initial segregation of the dataset for the eight states yielded 1,202,870 observations with 972, 521(81%) being individual entities and 230, 349 (19%) being non-individual entities. Further filtering showed the subset of individual entities to be used in the analysis consisting of 647,957(67%) females and 324,564(33%) males.

Figure 1: Gender distribution for eight states

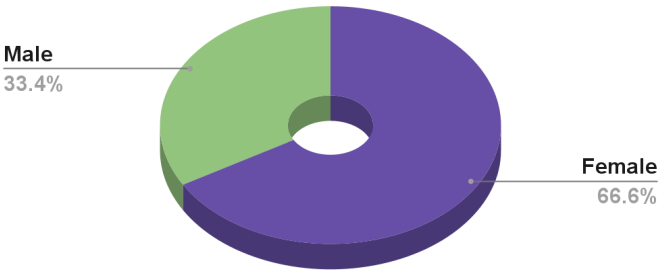
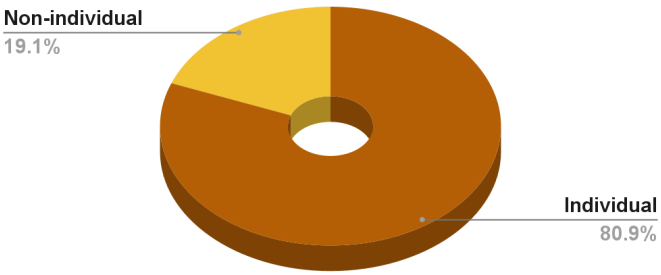


Figure 2: Distribution of entities for the eight states



The statistical analysis showed that there was a significant difference based on gender in the distribution of solo establishments in the eight states under observation ($p \leq 0.05$). Further observation showed that there were more females owning solo establishments (217,780) than there were males (107, 286) within Hawaii, Michigan, Minnesota, Mississippi, New York, Oklahoma, South Dakota and Tennessee altogether.

Part 3 - Gender Difference In Choice Of Speciality

Like the section before it, this section details the exploration of gender differences in the choice of specialty, specifically the low risk-low reward specialties and high risk - high reward specialties. The analysis tests the hypothesis: male doctors are more likely than their female peers to choose the practices that are associated with higher risk for a higher reward. For the purpose of this analysis, Obstetrics and Gynecology, and Pediatrics were defined as the low risk-low reward specialties, while Surgery and Orthopedic Surgery were defined as high risk - high reward specialties.

Method: The statistical software used for this analysis was Python. The taxonomy codes for the low-risk and high-risk specialties were pulled from the Healthcare Care Provider Taxonomy Code Set and can be seen in Table 2.

Low risk- Low Reward Category	Obstetrics & Gynecology	207V00000X
	Pediatrics	208000000X
High risk- High Reward Category	Surgery	208600000X
	Orthopaedic Pediatrics	207X00000X

Table 2: List of code of low and high risk-reward categories

The sub-dataset containing data for the eight states this report focuses on was then used to analyze the categories by gender. Once the 2 x 2 contingency table was established, a Fisher's test was run to determine whether gender had any effect on the choice of low-risk or high-risk specialties.

Result and Analysis: Using Fisher's exact test, a highly significant difference was established ($p \leq 0.001$). In light of this, the null hypothesis that gender is independent of specialty preference can be rejected within the eight states - Hawaii, Michigan, Minnesota, Mississippi, New York, Oklahoma, South Dakota and Tennessee. The data shows sufficient evidence that a doctor's gender does in fact affect their choice of specialty. The result as displayed in Table 3 indicates that, a greater proportion of male doctors prefer high risk and high reward specialty, and this supports the original hypothesis that male doctors are more likely than their female peers to choose the practices that are associated with higher risk for a higher reward.

Gender	High Risk-Reward	Low Risk-Reward
Females	1624	11839
Males	8247	7192

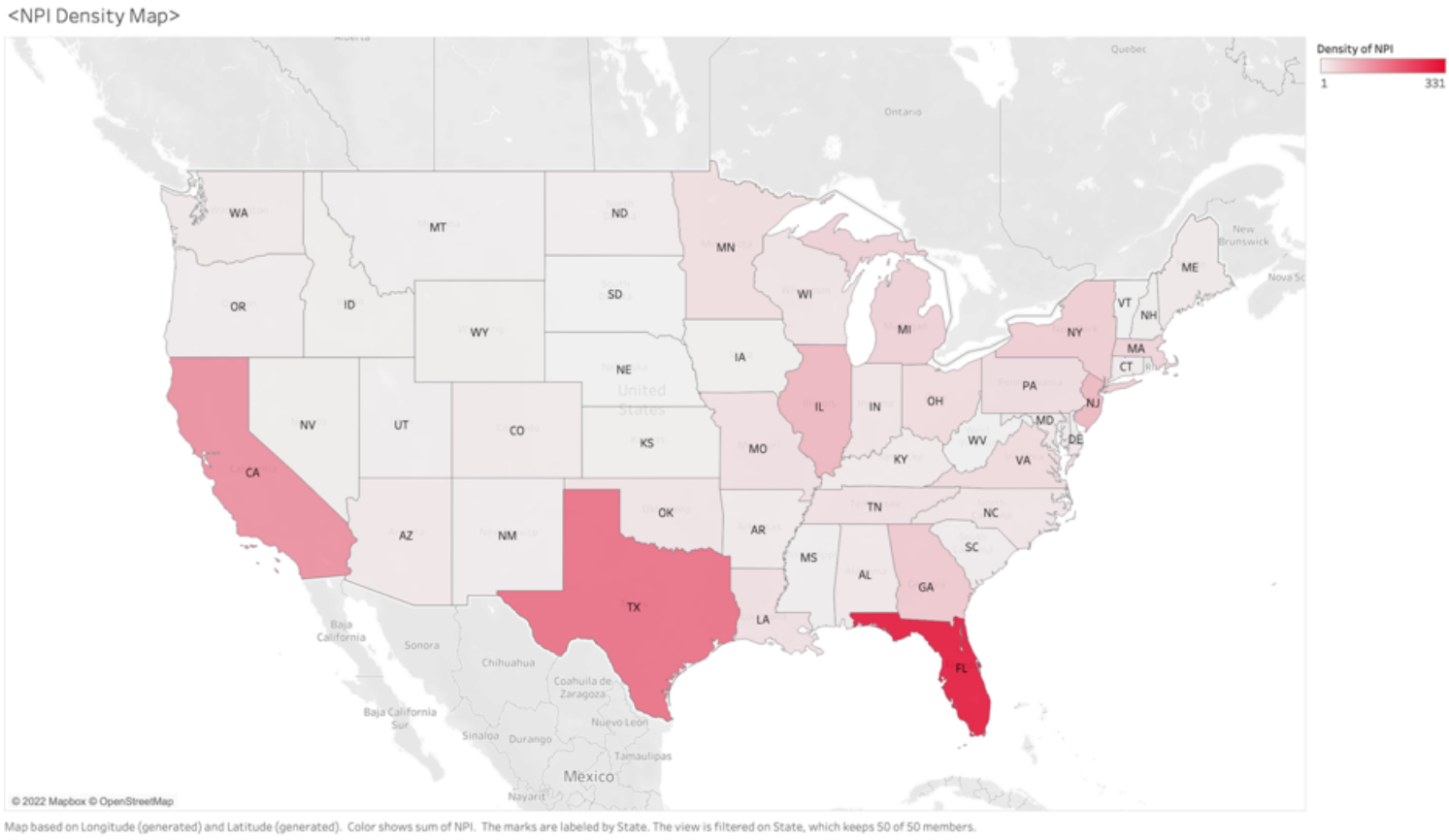
Table 3: Male and female providers by low and high risk-reward categories

Part 4 - National Distribution of MRI centers

Using the complete database, this portion of the analysis sought to explore the national distribution of MRI centers across the United States relative to the population of the respective states.

Method: All individual entities were filtered out of the main dataset to create a subset consisting mainly of healthcare facilities. The subset was further filtered to eliminate all non-MRI facilities using the taxonomy code "261QM1200X". A summary of the number of MRI centers in each state was carried out and population data was obtained from the U.S. Census Bureau with estimated figures on July 1, 2020. The MRI density was then calculated by dividing the number of MRI centers by the population figures.

The heat map represented by the MRI density of each state is shown in figure 3. The denser colors indicate that the state has a higher MRI capacity in terms of the number of MRI centers per 1000,000 population.



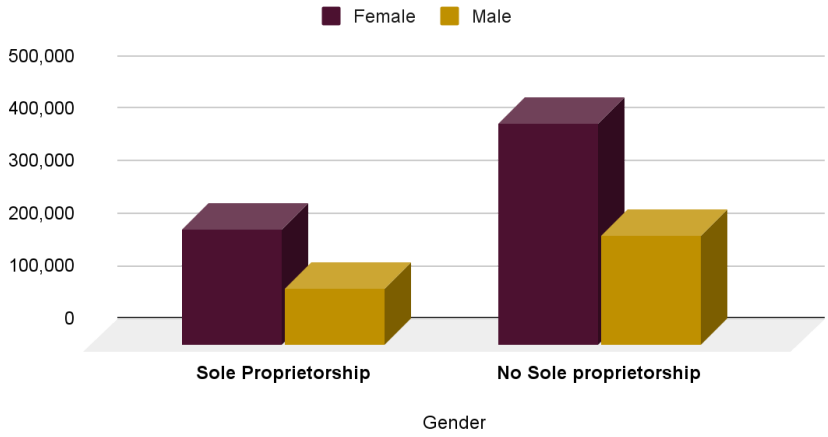
Map based on Longitude (generated) and Latitude (generated). Color shows sum of NPI. The marks are labeled by State. The view is filtered on State, which keeps 50 of 50 members.

Figure 3: Heat map showing the national distribution of MRI Centres across the states of the United States of America

Discussion and Limitations

From the series of analysis carried on the subset of the data relevant to this report, i.e., for the eight states, Hawaii, Michigan, Minnesota, Mississippi, New York, Oklahoma, South Dakota and Tennessee, the following observations were made. There was a statistically significant difference in the distribution of sole proprietorship between individuals in the NPPES based on gender in the eight states. The analysis revealed that more females had sole establishments than the men as displayed in figure 4.

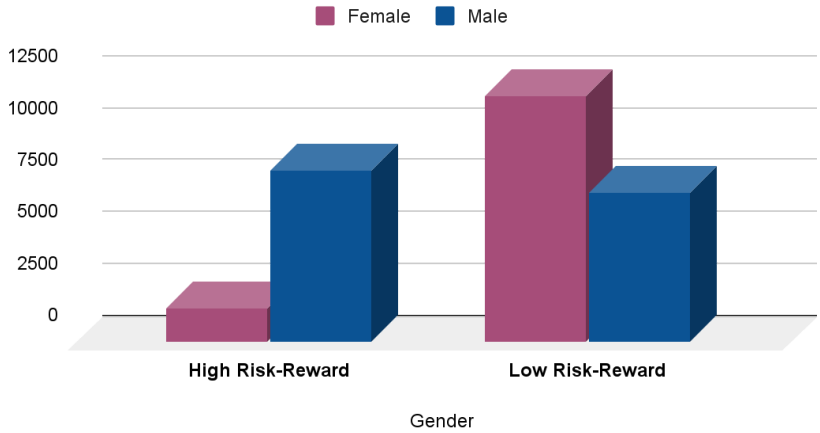
Figure 4: Distribution of sole proprietorship based on gender



It can be established from the data that generally in the eight states, there are healthcare providers in paid employment than there are owning solo establishments. Taking into account the fact that there are two times more females than males in the states being observed, one could argue that there would naturally be more females on both sides of the divide as seen in figure 4. However, the data shows more females having sole proprietorship in the eight states than their male counterparts. Also taken into consideration is the fact that across the Unites States, there has been a steady rise in the number of women becoming physicians since 2007 (Boyle, 2021).

Looking at the analysis on specialities, the hypothesis that male doctors are more likely than their female peers to choose the practices that are associated with higher risk for a higher reward was confirmed ($p \leq 0.001$). This result may be attributed to the fact that the higher risk specialties have more rigorous and time-consuming trainings than most others. A piece by Patrick Boyle (2021) cited Pediatrics and Obstetrics and Gynecology as having the highest and second highest percentages of women, with orthpedic surgery and other types of surgical specialties being male-dominated. This aligns with the results obtained from this analysis though carried out on the data for eight states.

Figure 5: High Risk-Reward and Low Risk-Reward



It is worth noting that the idea that the male-dominated specialties are more physically

demanding seem to be the commonly cited point or argument when issues regarding gender imbalance among specialties are raised.

Moving on to the analysis on the distribution of MRI centres, clearly indicated was the fact that Texas, Florida and California, states with the most intense colors, have the highest MRI density across board, Florida being the most intense of the three. WIth all these states being located in the southern region, one possible explanation could be the fact that people living in the warmer climate have a higher chance to develop symptoms requiring the use of MRIs.

One limitiation worth mentioning however is that though the data set was representative of the number of MRI facilities in the various states, there was insufficient information as to the number of MRI machines each facility had. Therefore, this may not be the most accurate representation of the national distribution.

Conclusion

From the analysis carried out using various subsets of the most recently updated NPPES dataset, we see that there are clear gender imbalances along the lines of sole proprietorship and choice of specialities, with more females being predisposed to having their private establishments in the eight states this report observed, and females being pre-disposed to lower risk specialties. This report also helped to visually map out the national distribution of MRIs in the United States showing certain southern parts like Florida having more capacity that other states.

References

NPPES. (n.d.). NPPES. Retrieved January 29, 2022, from https://nppes.cms.hhs.gov/#/

Medicare Learning Network. (2021). *NPI: What You Need To Know* (Vol. MLN909434). https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/NPI-What-You-Need-To-Know.pdf

Boyle, P. (2021, February 2). *Nation's physician workforce evolves: more women, a bit older, and toward different specialties*. AAMC. Retrieved January 30, 2022, from https://www.aamc.org/news-insights/nation-s-physician-workforce-evolves-more-women-bit-older-and-toward-different-specialties

Appendix 1 - Code for Part 1

```
import pandas as pd
df1 = pd.read_csv("npidata_pfile_20050523-20220109.csv",
                  usecols=['NPI','Provider License Number State Code_1',
                  'Provider Last Name (Legal Name)','Provider First Name'])
df1.query("NPI=='1972507325'")
```

Appendix 2 - Code for Part 2

```
//import data (csv)//
import delimited "/Volumes/Bloomn'💜/HDADM Tools/NPPES_Data_Dissemination_January_2022/npidata_pfile_20050523-20220109.csv"
//save as .dta file//
save npidata20220109.dta, replace
//filter for group's states//
keep if v32 == "HI" | v32 == "MI" | v32 == "MN" | v32 == "MS" | v32 == "NY" | v32 == "OK" | v32 == "SD" | v32 == "TN"
//save cropped grp data//
save npidata20220109_grp4.dta, replace
//drop non-individual entities//
drop if entitytypecode ==2
//view gender + missing//
tab providergendercode, missing
//view missing in sole prop. var//
tab issoleproprietor, missing
//drop values other than Y or N//
drop if issoleproprietor == "X"
//Fisher's (chi2)//
tab providergendercode issoleproprietor, chi2 exact
```

Appendix 3 - Code for Part 3

```
import pandas as pd
import scipy.stats as stats
import numpy as np
df3 = pd.read_csv("npidata_pfile_20050523-20220109.csv",
                  usecols=['NPI','Provider Business Practice Location Address State Name',
                       'Provider Gender Code','Healthcare Provider Taxonomy Code_1'])
df3.head()
states = ['HI','MI','MN','MS','NY','OK','SD','TN']
df3_state = df3[df3['Provider Business Practice Location Address State Name'].isin(states)]
df3_state.head()
categories = ['207V00000X','208000000X','208600000X','207X00000X']
df3_risk = df3_state[df3_state['Healthcare Provider Taxonomy Code_1'].isin(categories)]
df3_risk.head()
df3_FH = ((df3_risk['Provider Gender Code']=='F') &
        ((df3_risk['Healthcare Provider Taxonomy Code_1'] == '208600000X')|
         (df3_risk['Healthcare Provider Taxonomy Code_1'] == '207X00000X')))
df3_FL = ((df3_risk['Provider Gender Code']=='F') &
        ((df3_risk['Healthcare Provider Taxonomy Code_1'] == '207V00000X')|
         (df3_risk['Healthcare Provider Taxonomy Code_1'] == '208000000X')))
df3_MH = ((df3_risk['Provider Gender Code']=='M') &
        ((df3_risk['Healthcare Provider Taxonomy Code_1'] == '208600000X')|
         (df3_risk['Healthcare Provider Taxonomy Code_1'] == '207X00000X'))
df3_ML = ((df3_risk['Provider Gender Code']=='M') &
        ((df3_risk['Healthcare Provider Taxonomy Code_1'] == '207V00000X')|
         (df3_risk['Healthcare Provider Taxonomy Code_1'] == '208000000X')))
gr = np.array([[df3_FH.sum(),df3_FL.sum()],[df3_MH.sum(),df3_ML.sum()]])
df3_f = pd.DataFrame(gr, columns=['High Risk','Low Risk'])
df3_f.index = ['Females','Males']
df3_f
oddscratio, pvalue =stats.fisher_exact(df3_f)
Pvalue
```

Appendix 4 - Code for Part 4

```
import pandas as pd
npidata_4 = pd.read_csv("/Users/cynding/Desktop/Brandeis 2022 Spring/HS 256F - Healthcare Data Analytics/HW1/NPPES_Data_Dissemination_January_2022/npidata_pfile_20050523-20220109.csv",usecols=['NPI','Entity Type Code','Healthcare Provider Taxonomy Code_1','Provider Business Mailing Address State Name'])
npidata_4 = npidata_4[(npidata_4['Entity Type Code']==2) & (npidata_4['Healthcare Provider Taxonomy Code_1']=='261QM1200X')]
mri = npidata_4.groupby(npidata_4['Provider Business Mailing Address State Name']).count()
mri = mri[['NPI']]
mri.rename_axis('State',inplace=True)
population = pd.read_excel('/Users/cynding/Desktop/Brandeis 2022 Spring/HS 256F - Healthcare Data Analytics/HW1/Population.xlsx')
population.head()
population['population']=population['July 1 2020']/1000000
#population.set_index(['State'])
df = mri.merge(population.drop(columns=['Geographical Area','July 1 2020']),on='State')
df['NPI Density']=df['NPI']/df['population']
df.to_csv('NPI Density.csv')
```