Graduate Project Writeup

Nianxu Wang

December 13, 2021

Introduction & Motivation

In this project we are provided with data pertaining to COVID-19 and health care.

I am interested in delving deeper into the state of mental health during the pandemic, and pinpoint what variables are affecting it given the data on age, state of residence, gender, ethnicity, as well as situations like insurance availability and healthcare access. My intuition is that a younger population are less likely to have stable access to healthcare and insurance, and therefore a higher rate of anxiety and depression. For this I will explore trends that link age-based data on insurance rates and healthcare access with self-reported mental health symptoms data, while being careful of the rules of correlation and causality.

From the following statistical analyses, my goal is to understand the state of mental healthcare in the United States, and reasons for its current distribution. The pulse surveys provided to us is only a tiny random sample of the population of the U.S. but serves well enough in helping us draw conclusions on the trends in people's mental health and access to healthcare, which helps describe our population parameters.

Survey of Related Works

It's been almost two years that the COVID-19 pandemic has raged on, and there is an abundance of data science work done by governments, companies, and global organizations everywhere due to the current relevance and pressing concerns. One highlight is the work done in our very own class, where COVID-19 datasets are options for both the final and graduate project. Specifically comparing with the final project, it was concerned with researching the number of cases directly, as well as straightforward prevention methods like vaccinations and mask usage. The topics for my project focus on periphery issues such as insurance and access to healthcare, and most importantly, mental health, which is often overlooked. A web search for COVID and mental health reveals many news articles, but not as many scholarly research and a dearth of available data, with one paper I found in fact referencing the very same NCHS datasets we are using. Overall, I believe there's not much existing research exploring mental health impacts of COVID, so hopefully our graduate project can present a new perspective on this pandemic.

External Datasets

When perusing through documentation for the anxiety and depression dataset, they suggested a comparison with a similar 2019 dataset.[4] The major differences are that this 2019 dataset aggregated all adults over the age of 18 (rather than split by age groups like our data), and the split is per month, rather than the 28-ish arbitrary time periods. Due to the time mismatch, we will primarily reference the average value for 2019 provided in this new dataset as a pre-pandemic benchmark when looking at the state of mental health during 2020.

Methodology

Causal Inference:

Inference is about using statistical techniques to analyze our data and draw conclusions about the trends and relationships between variables. I loved the analogy in class regarding house prices. Typical predictive modeling would simply tell us what our prediction of the price is given a set of features of the house. Inference is thinking about the impact of individual features, for example, how much extra value does an extra garage spot add to a house. When it comes to causal inference, we are looking at the difference between correlation and causation. For example, is it really the extra garage spot adding value to the house, or is it because that increases the overall square feet area of the house, indirectly raising the price? Data alone is not enough to tell us about the entire relationship, and if we take a blinkered approach by only looking at data, we risk falling into the trap of mistaking coincidental correlation for causational effects. The best way to reduce causal and correlation mistakes is to be or have a domain expert in that field because the knowledge allows us to make relevant decisions. I consider myself well read in current events and generally informed, so I should be able to make common sense assumptions that lead to reasonable conclusions about datasets involving healthcare access and insurance. However, I do not claim to be an expert in mental health, a limitation that I acknowledge when exploring datasets involving anxiety, depression, and therapy.

Evaluation of Approach & Method Limitations:

Because I am using COVID-19 Dataset B for my project, there is no modeling requirement. The nature of the datasets we were given prevents the use of any predictive modeling we learnt in class. First, it may seem like we have thousands of rows in each dataset, but that is an illusion, because the data is split by indicators, and already aggregated into subgroups like ages or by state. The granularity of each datapoint is the percentage of a surveyed subgroup who answered yes to that indicator in that time period.

Using techniques such as bootstrapping to create multiple models is very difficult. We cannot sample with replacement to bootstrap any training data, because we only have a single data point for a particular subgroup in one testing period. While we do have multiple testing periods for each subgroup, if we're looking to resample with replacement, that is not an effective decision because the sample size is very small, and that would ignore how the data changes over time. Plus, there are at most 28 time periods, with some dataset like the mental counseling one having even fewer, making time series or any modeling difficult. Therefore, each logical group of datapoints is tiny. Another problem is that the surveys reinvited the same participants back for future questionnaire phases. This is not independent but can be good if we want to observe a more consistent group of people for changes over time. Overall, this means when doing analysis and trying to infer conclusions, we need to use these pre-aggregated data as is, and to do more time-based analysis.

My approach to dealing with time periods is to logically pivot into subgroups and simply plot them out, which works surprisingly well because there's no overplotting with so few datapoints. The trends are clear visually. With categorical data, since everything is neatly aggregated for us, I use barplots and heatmaps as appropriate.

I also make cautious choices about what indicators in the mental health-related datasets to use, for example by not differentiating between anxiety and depressive disorder symptoms, and instead using the combined indicator of both. The reason is I do not claim to be an expert enough to differentiate between them; I consider them both to be important, and it was convenient for graphing.

I have dealt with missing data and acknowledge that there are biases and issues with the data collection. Detailed notes regarding these and other analysis decisions can be found in my analysis notebook, which I leave out here to prevent it being too convoluted.

We can see that since the dataset provided to us are already aggregated data, they also supplied us with the 95% confidence intervals, meaning we do not have to bootstrap sample any confidence intervals from raw data ourselves. The value column that exists in the four NCHS datasets is the center of the confidence interval and is the mean.

Exploratory Data Analysis, Inference & Analysis of Results

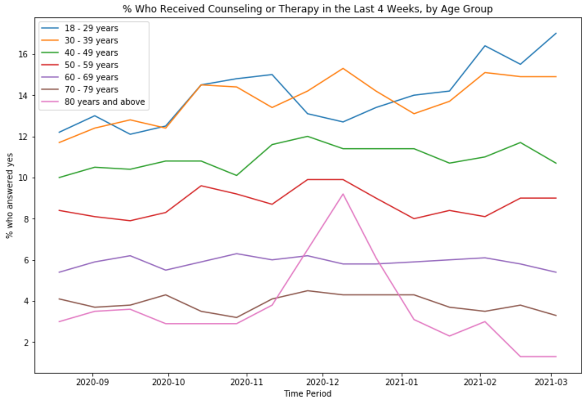

Figure 1

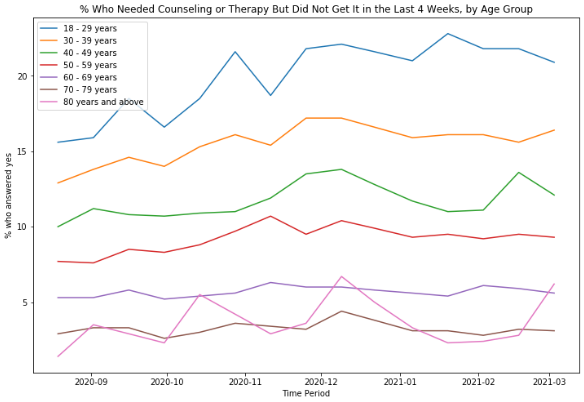

Figure 2

In figure 1 and 2, we can see a clear stratification between the age groups, where the older the group, the less likely they are to seek counseling or therapy. The reason shouldn't be financial, if anything the younger generation tends to have less money saved, but rather awareness and perspective. Conversations around therapy and mental health has gotten louder in the past decade or so, especially amongst millennials and gen z, but the older generations aren't as aware of the need for therapy.

Another noticeable trend is that for the 18-29 and 30-39 age groups, the rate is gradually increasing as the pandemic goes on, starting with around 12% in August 2020, to around 16% this March. In contrast the older generations are staying flat or even decreasing. There is one spike for one month for the over 80 age group, rising from 11-25, peaking in 12-09, and decreasing again in the time period 12-22. This hill is an interesting spike, as no other age group has this spike. I want to think this may be because the older generation, when getting closer to Christmas time, are missing their families (due to restrictions it's hard to see anyone), and therefore seeking therapy. However, then there should be similar spikes for the 70-79 and 60-69 age ranges too. More likely this is an outlier, or more accurately, something going wrong in the data collection process, as the pulse survey data is not super reliable. The 80+ age group data also had quite a few missing values. Later in the societal impacts section, I talk about the lack of representation for the older generation are not well represented enough.

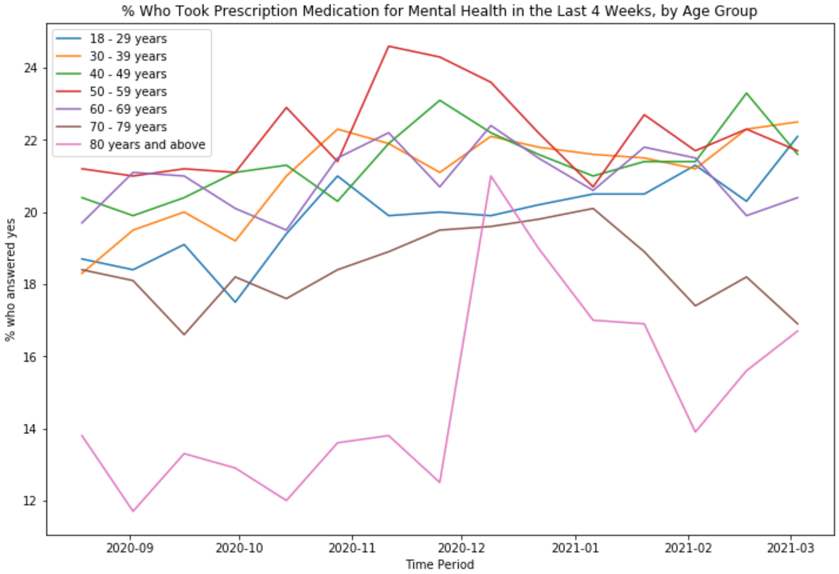

Figure 3

I had expected figure 3 to look like figures 1 & 2. There are certain similarities such as the older two generations tend to have lower percentages. The stratification isn't as clear-cut as before, namely the groups older than millennials (40+) are catching up in the rate of prescription mental health medicine usage. I also noticed that the percentages are consistently higher than those who sought therapy for every age group. This could mean many things: one is that people are self-diagnosing and taking medicine. Also given rising acceptance of drugs such as Marijuana, which is also often associated with mental health, it helps explain why the percentages are higher, given that while you still need prescriptions in many cases, you don't need to necessarily see a mental health specialist or therapist.

Doubtful, I decided to look back at the original questions asked in the surveys, and I believe I found the most relevant question (given there were no other similar questions):
"Q38a) At any time in the last 4 weeks, did you take prescription medication to help you with any emotions or with your concentration, behavior or mental health?"[5]

This question is more than just mental health, but also behavior and concentration. This is a much wider net to cast, plus we also need to think about how people interpret this question, so responders may have taken other medicines in account. This also helps explain why there's a consistently higher percentage for every age group, because technically this statistic encompasses more than medicine strictly prescribed for mental health after seeing a therapist.

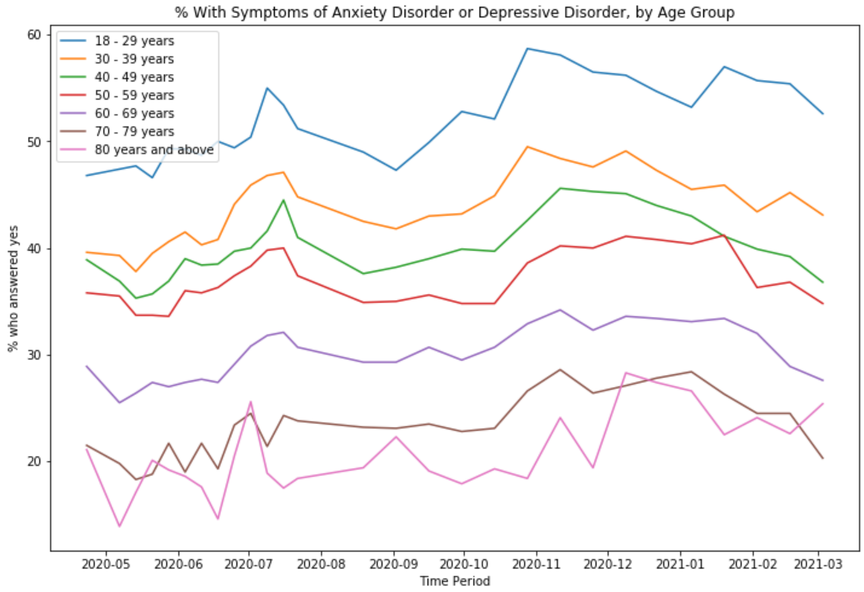

Figure 4

There is a major similarity between this plot and figures 1 & 2, namely the clear stratification and order from young to old generations where the younger you are the more likely you are to experience anxiety and depressive symptoms. This reinforces the idea that there is a lack of awareness for mental health amongst the older generation, partly because growing up a few decades ago, it wasn't part of the social conversation.

Across the board we see that the percentage who has symptoms is a lot higher than those who want to seek help. Therefore, there needs to be more awareness for everyone. Maybe people don't want to acknowledge such a sensitive issue, and if so, there could be more availability of private and confidential therapies. Mentioned in the analysis notebook, when exploring the mental health therapy dataset, I talked about how there is a significant percentage of people who needed counseling but could not get access to any. The data shown in figure 3 further proves the point that we need more health infrastructure and policies to deal with a rising mental health crisis.

One major reason for younger generations having more depressive symptoms as well as likelier to seek therapy is the economic impact of COVID, especially job loss. There have been many layoffs through the pandemic, which affects younger people more, because older people are either retired or have more savings. Other past research as well as common knowledge tells us that loss of employment and economic bear markets causes more symptoms of anxiety and depression, and this is exactly what the pandemic has done. Finally, the 18-29-year-old age group also has other issues to contend with, such as closures of universities and other social venues (such as bars and clubs) that young people hang out in. The lack of social interaction and loneliness is also causing a much higher rate of depression, a widely accepted fact.

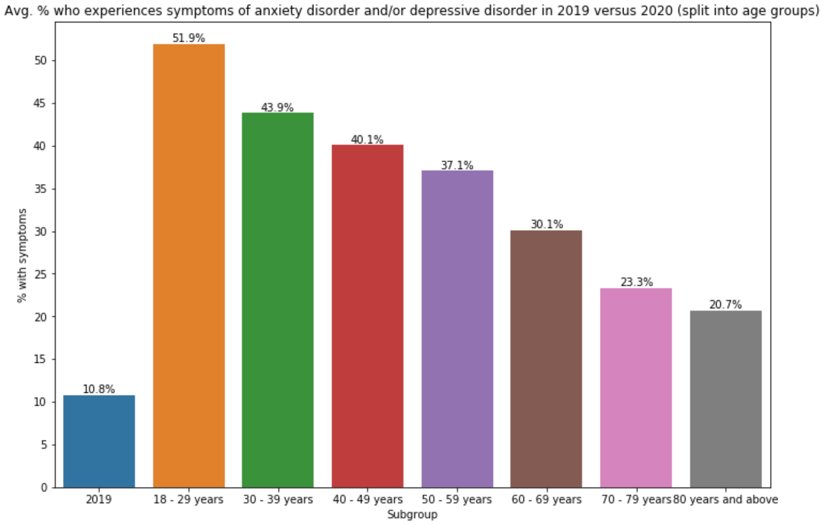

Figure 5

When I was reading through the documentation, there was mention of a 2019 NCHS dataset that dealt with anxiety and depressive symptoms.[4]

Here we have plotted the average percentage of adults (18+) who experiences symptoms of anxiety disorder and/or depressive disorder in 2019 and compared with the average percentage per age group for the year April 2020 to March 2021.

If we look over the table for all 12 months of depressive symptoms data in 2019 (available in analysis notebook or reference link [4]), we can see that the percentages are consistent hovering around 10-11%, with the average being 10.8%. But even the smallest average percentage in 2020, the 80 years and above age group, is twice that value. During the pandemic year, there has been a drastic increase in those who need mental health counseling more than ever, and with all the stress from reasons noted previously, this should be a priority focus with government policy.

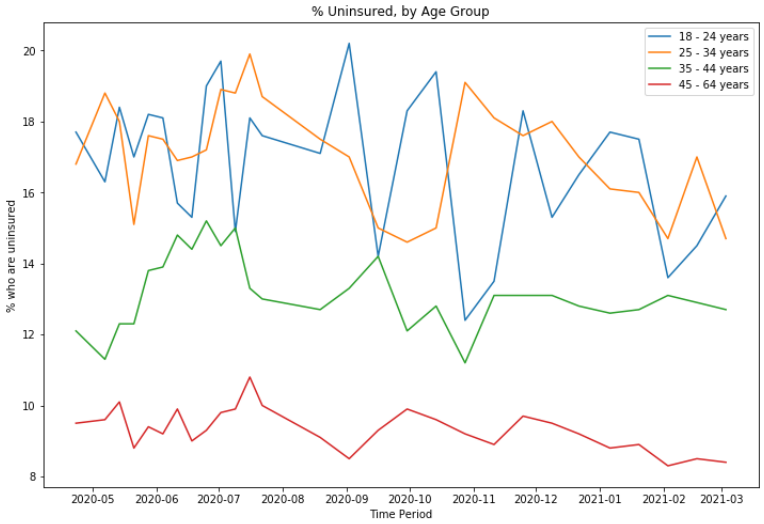

Figure 6

For figure 6, which deals with the uninsured, the first major issue we see is that unlike the mental and depression datasets, the insurance dataset uses a very different age group cutoff, which makes direct comparison with the previous two datasets very difficult. These data come from the same surveys, so I am unsure why they chose to pick different age groups. Nevertheless, this can help inform us of general trends between generations.

The reason for the large fluctuations amongst young adults and millennials could be attributed to uncertainties in employment and school, two major sources of health insurance. Small periodic stimulus checks may help some people, but it does not replace steady income, and could explain the almost monthly up and downs for health insurance coverage.

In contrast, older gen X and boomers have lower and more consistent uninsured rates, indicating more savings and less likely to be perturbed by the economic disturbances of COVID. Another reason could also be fewer things in life requiring money, so they could simply be resting in care homes and have most of the remainder of their life sorted, plus older people tend to maintain more permanent health insurances that cover them until end of life. On the other hand, younger generations may have families to take care of and directly spend money on, and have more demands, such as for entertainment, plus youthful confidence / over-confidence means they may not place health insurance as a priority for spending. One suggestion to improve healthcare access could be to encourage buying insurance, which as we all know is extremely important given the state of healthcare in the U.S.

Exploring data by state: Symptoms and counseling¶

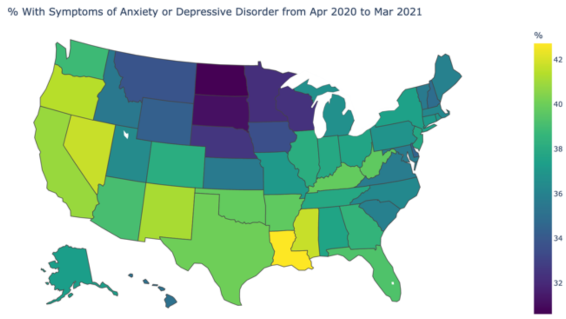
% With Symptoms of Anxiety or Depressive Disorder from Apr 2020 to Mar 2021
%
42
40
38
36
34
32

Figure 7

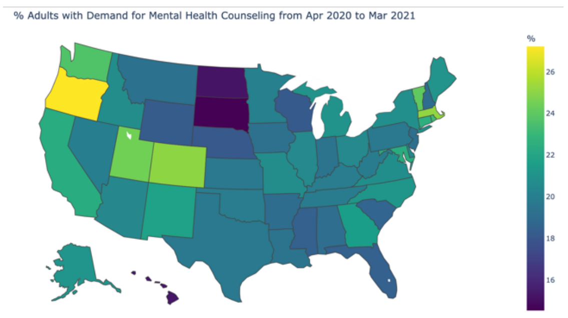
% Adults with Demand for Mental Health Counseling from Apr 2020 to Mar 2021
%
26
24
22
20
18
16

Figure 8

There are some similarities for the states around central north America, notably North and South Dakota and the surrounding states, which appears to have generally little demand for therapy, as well as lower displays of symptoms of anxiety or depressive disorder.

Part of the reason could be because there's not a lot of people there: North and South Dakota, and Wyoming next to them, are amongst the least populated states, beaten only by the likes of Alaska and District of Columbia. As for why fewer people results in fewer reported symptoms and less seeking therapy, it could be a variety of reasons: maybe a lack of educational awareness (these few states are also among the poorest), maybe there's not enough resources to afford privacy on sensitive topics, especially in a lot of small towns where everyone knows each other.

In terms of politics, the distinction isn't quite clear between red and blue states and how that may relate to mental health. Looking at the map for symptoms, while we can see a few blue states with high rates of symptoms (e.g. CA, OR, NV), there's also plenty of red states with high demand too (e.g. TX, LA, MS). Looking at the demand for mental health counseling map, we can see lower demand across the board. This is like when we looked at age groups, that symptoms of depression tend to be higher than those who want to seek help, showing that we may need to increase awareness of the importance of mental health. Again, the percentage of demand for mental health counseling does not seem to clearly define any political boundaries, the demand is fairly similar across all the states, hovering from 20 to 30%.

Healthcare access and insurance by state:

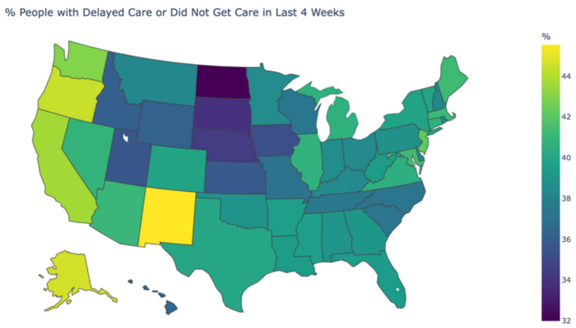
% People with Delayed Care or Did Not Get Care in Last 4 Weeks
%
44
42
40
38
36
34
32

Figure 9

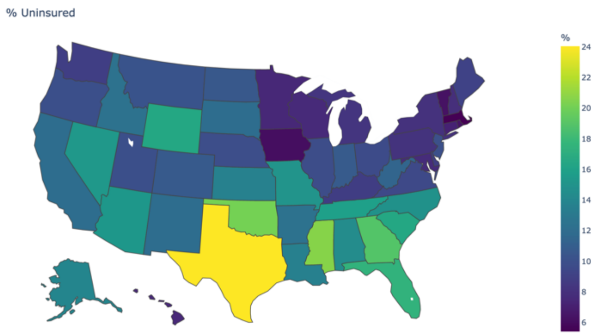

Figure 10

Just looking at figure 8, we can also see that there's not much of a clear political divide. Rather, it appears to be a population issue, i.e., states with high populations tend to have a higher percentage with not getting care. Some of the least populated states mentioned before, North/South Dakota...etc. has the lowest percentages, whereas highly populated states like California, New Mexico...etc. are the opposite. This may be because the more populated states have bigger cities, with medical facilities already strained by the pandemic, there isn't enough resources to go around. Not to mention that closely packed cities are far bigger spreaders of COVID and diseases.

The insurance map is interesting, because we can see that Texas sticks out like a sore thumb. A quick google search reveals that Texas is in fact the uninsured capital of the U.S. [6] However interestingly enough, this does not seem to affect Texas too much in terms of the percentage of people with lack of access to healthcare. Of course, this doesn’t mean access isn’t an issue in Texas, a large percentage of people without access may fall into the no insurance group. It’s hard to tell from the provided data alone, which is why more research is needed into the single problem.

Finally, there does not seem to be a major correlation between a state's insurance rate and their access to healthcare. This goes against my earlier remark that maybe encouraging more people to take on insurances will help their access to healthcare. The situation with healthcare is complicated and dependent on individuals.

Correlation between depressive symptoms and lack of healthcare access:

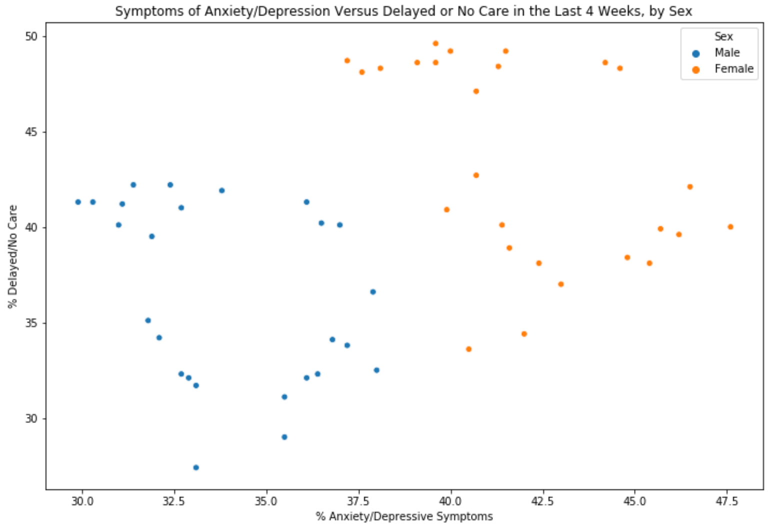

Figure 11

This is a very interesting result showing that across the entire time period for the surveys, females consistently reported higher rate of anxiety and depressive symptoms. And although on average females also experienced a higher rate of delayed care, the percentage of people not getting care if worryingly high regardless of sex. This discovery may help inform certain policymaking, such as emphasis on female mental health care advertising and initiatives. Of course, there are a few caveats we need to consider. We are often taught that men don't tend to reveal emotions as much, and regardless of whether this is true or not, this is how the larger society sees it, and so the numbers for depressive symptoms for men may be higher than the survey admits.

Societal Impacts & Ethical Concerns

We need to acknowledge that there are groups of left out of this conversation. As I researched the details of data collection, I found a few surprising facts about the way the data is constituted and made me think about its limitations. The sampling frame of the NCHS data is people with emails or cell phones associated with a housing unit in the Census Bureau Master Address File Data. The way the data is collected is by emailing or texting the selected participants an internet questionnaire. There is a slight mismatch between the sampling frame and the population of interest, which should be the entire population of the U.S.

One instance whose impact of the pandemic we need to consider are the homeless, an important part of the social justice discussion here at Berkeley. Because they have trouble finding a permanent home, they are not included and represented in the survey. Arguably they are amongst some of the most affected people by the pandemic, and in dire need of mental and general healthcare.

Another notable group that is possibly left out are retirees living in retirement homes. Many older people may not have emails or even cell phones (just a landline), and so cannot respond to the questionnaire. We are also unsure whether retirement homes are strictly considered housing units in this survey. There have been multiple news reports about the terrible conditions inside retirement homes throughout the pandemic, in addition to how disproportionately affected old people are by the disease. This is partially evident through the fact that there tends to be more missing values for the 80 and older age group in the datasets. That made the 80+ age group data less reliable, because we don’t have many datapoints in the first place. However, I do want to acknowledge that the Telemedicine survey does try to compensate for the lack of data on over 65. In the Telemedicine dataset, approximately 27.3% of the samples are for over 65, which is calculated by taking the number of samples in the Subgroup “65 years and over” and dividing by the total in “Age groups.” This contrasts with U.S. demographics where around 16.6% are over 65 years of age. [1] Despite the dataset supposed being calibrated to U.S. population counts, I believe the discrepancy is due to two reasons. First, the data collection is outsourced to commercial vendors, with participants often accepting payment or prizes. The CDC acknowledges that this type of survey may have more bias and less accuracy than traditional methods.[2] Secondly, this may be a result of the vendor’s implemented weighting and estimation method to account for nonresponse and under coverage of certain groups.[2]

I delved deeply into the technical documentation of the surveys. Looking at the coverage rates from the Source and Accuracy Statements for the NCHS, there were a lot of undercoverage for various groups. First, the 65+ age group was a little underrepresented. Then the 18-24-year-old group too, because many may not have a set household registered to the census bureau yet, due to school or not making enough during early career. Low educational groups such as those without a high school diploma are also underrepresented, with possible reasons such as a lack of registered housing, or difficulty in gaining access to internet due to poverty.[3]

Finally, another sizeable group that are left out of the Telemedicine datasets are non-English speakers and certain groups of immigrants because the survey was conducted exclusively in English. Undocumented immigrants may also not be present in the databases for survey. Overall, this pulse survey fails to properly capture data from the most vulnerable and marginalized members of society.

Conclusion & Future Work

In the future, I would hope the NCHS being a federal agency would be more inclusive and recognizing of a variety of people from all walks of life. In turn, the comprehensive data collected would be for the benefit of everyone, because if we can understand these high-risk groups, we can reduce the rate of COVID-19 overall. A better ability to understand the mental state of people can help inform government policies and measure the effect of the significant changes the pandemic has wrought on everyday life. To do this, it would be best if we were able to discuss with domain experts in COVID and social impact.

For myself, a future step cold be taking a closer look at other indicators and relevant factors. I primarily looked at ages and state-wide data in investigating healthcare access and mental health. The notebook/writeup is already extremely long, but if it were longer, we can also determine whether there are any ethnic disparities and explore the role of gender in mental health more. If we are careful about using the educational attainment variables present currently, we may also look at how income affects mental health and access too, though we need to once again be careful to make certain conclusions that say fewer degrees equals more issues. These factors may be able to tell us more about social and cultural attitudes and perception towards mental health.

Overall, my analyses are good for very broad estimates over a representation of the entire American population.

References

[1] O'Neill, Aaron. “U.S.: Age Distribution.” *Statista*, 21 July 2021. https://www.statista.com/statistics/270000/age-distribution-in-the-united-states/.

[2] “Telemedicine - Research and Development Survey - COVID-19.” *Centers for Disease Control and Prevention*, 6 August 2021. https://www.cdc.gov/nchs/covid19/rands/telemedicine.htm

[3] “Household Pulse Survey Technical Documentation.” *United States Census Bureau*, revised 1 December 2021. https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html

[4] “Estimates of Mental Health Symptomatology, by Month of Interview: United States, 2019.” *National Center For Health Statistics*, Released March 2021. https://www.cdc.gov/nchs/data/nhis/mental-health-monthly-508.pdf

[5] “2020 COVID-19 Household Pulse Survey Phase 2.” https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/Phase_2_Questionnaire_11_2_20_Updated_English.pdf

[6] “The Uninsured in Texas**.**” https://www.texmed.org/uninsured_in_texas/