Analysis of effects of COVID-19 on the Mental Health of Population

Anisha Chazhoor (aschazho)

1 Introduction:

The COVID-19 pandemic has had a significant impact on the mental health of the global population. It has led to increased levels of anxiety and depression due to isolation, uncertainty, loss of loved ones, financial difficulties, and job loss. The aim of the project is to verify using a statistical approach, whether there has been any significant change in the mental health conditions of the population post-pandemic. Various tests such as McNemar and Wilcoxon Signed-Rank Tests have been performed to validate the hypothesis. Analysis using Chi-Square test has been performed to establish correlation between depression and other health factors. Causal analysis has been performed using Bayesian Networks with the correlated information, to understand cause-effect relationships. The code and output can be found here.

2 Literature Survey:

Artificial intelligence (AI) methods have been used to assist mental health providers, including psychiatrists and psychologists, for decision-making based on patients' historical data [1]. Models have been implemented to determine a subject's depression condition using EEG information [2], fMRI data [3], unstructured text notes in electronic medical records [4] and audio-visual records [5]. Meta-analysis of the COVID-19 pandemic, reveals that health care workers, non-infectious chronic disease patients, COVID-19 patients, and quarantined persons are at higher risk of depression, anxiety, distress, and insomnia [6].

3 Experiment and Methodology

3.1 About the Dataset

The National Health Interview Survey (NHIS) is an annual survey conducted by the U.S. National Centre for Health Statistics (NCHS) to collect information about the health and well-being of Americans. The NHIS collects data on a wide range of topics related to health care, including demographic information, health behaviours, chronic conditions, health care access and utilization, and health insurance coverage. We will be utilizing 3 datasets for the analysis: NHIS_2019, NHIS_2020 and NHIS_2020_Longitudinal. NHIS Longitudinal records the responses of surveys taken by the same individuals pre and post pandemic. NHIS_2019 contains 534 questions, NHIS_2020 has 617 questions. The combined dataset consists of 10415 participants, out of which 4790 are men and 5624 are women. There are many variables which record the mental health conditions of the participants. Some questions include history of depression diagnosis, medications, severity and frequency of depressive symptoms, anxiety levels and anxiety medications.

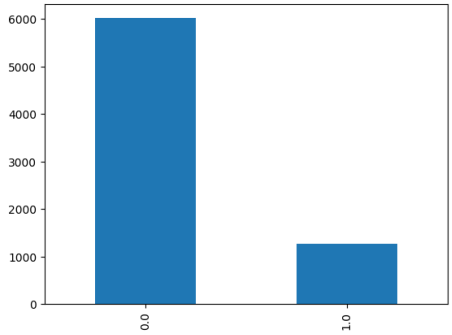

Figure 1: Class distribution of depression diagnosis variable

3.2 Data Pre-processing

The following steps have been implemented for data pre-processing:

- Merging NHIS_2020_Longitudinal with NHIS_2019/NHIS_2020 on unique household ID key: We will only be analysing the participants whose pre-post COVID-19 data is recorded.
- Manual dropping of columns: Removal of columns of object datatype. Removal of columns with 100% null values, as it hampers with data imputation. Some columns which have little to irrelevant in this particular use case, are also dropped.
- Identifying numerical and categorical data: Most categorical data are multi-labels, so convert them to indicator dummy variables using scikit-learn. Replacing “Refused”, “Not Ascertained”, “Don't Know” responses to Null.
- Normalize data: Numerical columns are normalized using RobustScaler() to avoid bias in the model.
- Handling missing values: MICE iterative imputer is implemented to handle missing values. The method involves creating multiple imputed datasets based on the existing data and using them to estimate missing values. Linear regression model and batchwise implementation is used, because of RAM limitations in Google Colab.

3.3 Exploratory Data Analysis

We will be using several EDA techniques to validate our hypothesis. Hypothesis tests to find correlation within the same year data are as follows:

- Chi-square test of Independence: The chi-square test of independence is a statistical test used to determine whether there is a significant association between two categorical variables. A contingency table is created with the observed frequencies and scipy.stats is used to perform the test.

$$H_0 = variable_i \; and \; \text{DEPEV\_A} \; (Depression \; variable) \; are \; indepdendent.$$
$$H_A = variable_i \; and \; \text{DEPEV\_A are not independent}$$

- Point Biserial Correlation: It is a measure of the strength and direction of the relationship between a continuous variable (numerical data in NHIS dataset) and a dichotomous variable (DEPEV_A).

$$H_0 = \; no \; significant \; correlation \; between \; var_i \; and \; DEPEV\_A$$
$$H_A = significant \; correlation \; between \; var_i \; and \; DEPEV\_A$$

For Pre-Post COVID-19 paired sample analysis:

- McNemar test: It is a statistical test used to analyse paired nominal data, where the same individuals are observed at two different time points. Here we will use it to understand if there is any significant change in the DEPEV_A and ANXEV_A (depression and anxiety respectively) metrics, between pre and post COVID-19 years. Implemented using statsmodels.stats, with $\alpha = 0.05$.

$$H_0 = no \; significant \; difference \; in \; the \; frequency \; of$$
$$depression \; and \; anxiety \; for \; pre - post \; data \; responses$$
$$H_A = significant \; difference \; in \; the \; frequency \; of$$
$$depression \; and \; anxiety \; for \; pre - post \; data \; responses$$

- Wilcoxon Signed-Rank Test: It is a non-parametric statistical test used to compare the means of two paired samples, measured on an ordinal scale. We will be using it to analyse the frequency of depression symptoms before and after the pandemic. The test is implemented

using scipy.stats, with $\alpha = 0.05$. Here, encoding of survey question is such that 1 represents maximum frequency and 5 the least. Hence, the hypothesis will be:

$$H_0\colon \mu_{2019} = \mu_{2020}\ for\ DEPFREQ\_A$$

$$H_0\colon \mu_{2019} > \mu_{2020}\ , i.e\ DEPFREQ_{2019} < DEPFREQ_{2020}$$

3.4 Causal Analysis using Bayesian Networks

Causal analysis can examine the factors that contribute to the development of mental health problems. Bayesian networks can be a useful tool for conducting causal analysis in mental health research. It is a probabilistic graphical model that represents a set of variables and the relationships between them. To use BN, we need to identify a set of variables that can contribute to depression and anxiety. Using chi-square test of independence, we get a list of variables which are co-related with the depression variable. We will study the relationships between DEPENV_A and some of these variables. BNLearn library has been used for structure learning. The Directed Acyclic Graph has been generated using exhaustive search, edge strength has been calculated using chi-square conditional independence test and non-significant edges are removed (pruning).

4. Results and Inference

The results for McNemar test for pre-post analysis of COVID-19 is shown in Table 1. For the variable depression diagnosis, the p-value is 0.015 < 0.05. Similarly for anxiety, the $p - value \approx 0$. Both indicate that there is a significant change in number of participants suffering from anxiety and depressive symptoms. The null hypothesis is rejected for both cases as p-value<0.05.

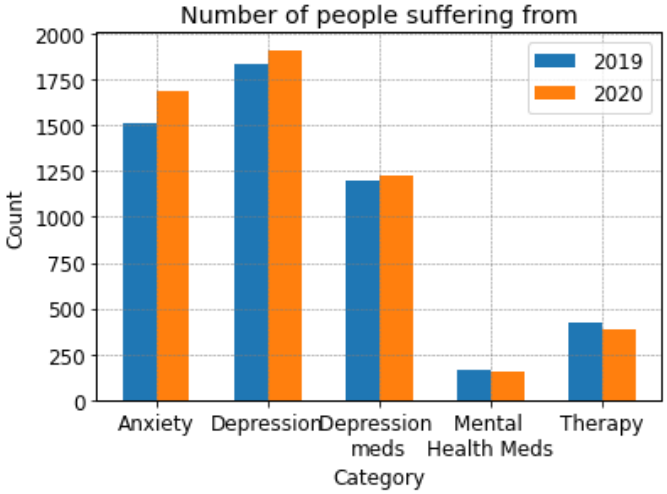

Figure 2: A plot of comparison between mental health related variables, pre-post COVID.

Variable	Test Statistic	P-value	Result
Depression	465.000	0.015255	Reject H_0
Anxiety	385.000	0.0000000068	Reject H_0
Depression medication	295.000	0.0742	Do not reject H_0
Anxiety medication	415.000	0.03505	Reject H_0

Table 1: Results of McNemar Test for paired Pre and Post COVID-19 data

Wilcoxon Signed-Rank Test for frequency of depressive and anxiety symptoms results in $p - values \approx 0$. Hence, the null hypothesis is rejected, as there is enough evidence to disprove it. This also affirms our alternative hypothesis that the frequency of depressive episodes has increased post pandemic.

Variable	Test Statistic	P-value	Result
Depression Symptoms Frequency	3814360.000	0.0000090054	Reject H_0
Anxiety Frequency	7921404.000	0.0000172204	Reject H_0

Table 2: Results of Wilcoxon Signed-Rank Test for paired Pre and Post COVID-19 data

For causal analysis, Bayesian networks are utilised. We will be using variables attributing to depression diagnosis (DEPEV_A), delayed counseling/therapy due to cost (MHTHDLY_A), extreme difficulty participating in social activities (SOCSCLPAR_A) and unable to work for health reasons/disabled (EMPRSNOWK_A) to understand the relationships and changes post COVID-19.

Constructing edges using chi-square test of independence, we get the table 3 for 2020 data and table 4 for 2019 data.

	source	target	stat_test	p_value	chi_square	dof
0	DEPEV_A_1.0	EMPRSNOWK_A_4.0	True	2.30176e-78	351.217	1
1	DEPEV_A_1.0	MHTHDLY_A_1.0	True	1.62414e-79	356.505	1
2	DEPEV_A_1.0	SOCSCLPAR_A_4.0	True	2.56134e-16	67.1149	1
3	EMPRSNOWK_A_4.0	SOCSCLPAR_A_4.0	True	8.92282e-47	206.275	1

Table 3: Compute edge strength with chi_square for 2020 NHIS data

	source	target	stat_test	p_value	chi_square	dof
0	DEPEV_A_1.0	EMPRSNOWK_A_4.0	True	9.77286e-110	495.349	1
1	DEPEV_A_1.0	MHTHDLY_A_1.0	True	4.8694e-99	446.189	1
2	EMPRSNOWK_A_4.0	SOCSCLPAR_A_4.0	True	2.63612e-47	208.702	1

Table 4: Compute edge strength with chi_square for 2019 NHIS data

Using the conditional chi-square table, we can observe that Depression was not previously associated with low social activity. Post pandemic, a new edge has been added between SOCSCLPAR_A and DEPENV_A. Thus, we can conclude that Depression is now associated with extreme social isolation. The graphical structure has been represented in figure 3.

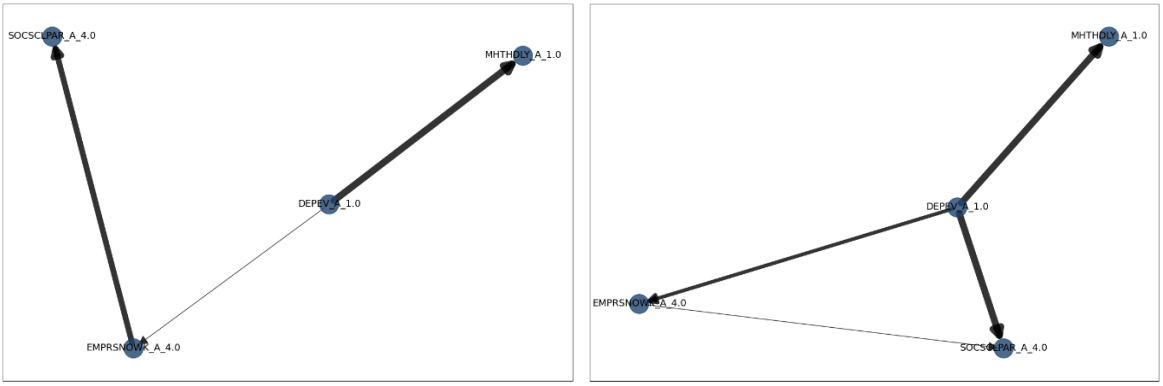

Figure 3: Comparison of Bayesian Networks DAG 2019 vs 2020

5. Conclusion and Future Work

We compared pre and post COVID-19 data using NHIS survey responses. Frequency of depressive and anxiety symptoms are compared and the hypothesis has been validated using McNemar Test. We got a p-value of 0.015255, thus concluding that there has been a significant change. Wilcoxon Signed-Rank test validated the hypothesis that the frequency has increased, with a p-value=0.0000090054. Causal analysis using Bayesian Networks confirmed a strong association between social isolation and depression after pandemic, using chi-square independence test. For further work, a more detailed causal analysis can provide insight about other factors affecting the mental health of participants. A post COVID-19 analysis with NHIS-2021 dataset can reveal the lasting effects of the pandemic on the mental state of the population.

6. References:

[1] Su, Chang, et al. "Deep learning in mental health outcome research: a scoping review." Translational Psychiatry 10.1 (2020): 116.

[2] Mohan, Yogeswaran, et al. "Artificial neural network for classification of depressive and normal in EEG." 2016 IEEE EMBS conference on biomedical engineering and sciences (IECBES). IEEE, 2016.

[3] Geng, Xiang-Fei, and Jun-Hai Xu. "Application of autoencoder in depression diagnosis." DEStech Trans Comput Sci Eng (csma) (2017): 146-151.

[4] Geraci, Joseph, et al. "Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression." BMJ Ment Health 20.3 (2017): 83-87.

[5] Valstar, Michel, et al. "Avec 2013: the continuous audio/visual emotion and depression recognition challenge." Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. 2013.

[6] Wu, Tianchen, et al. "Prevalence of mental health problems during the COVID-19 pandemic: A systematic review and meta-analysis." Journal of affective disorders 281 (2021): 91-98.