Research Proposal

A Comparative Analysis and Predicative Model for Reporting Depression in U.S. Adults

Kiran Mai Jaiswal Charpuria, Kruthika Gaddam, Mohith Surya Kiran Kasula, Bala Samantula, Sri Harsha Sudalagunta, April Taylor, Alan Varkey
[kichar, kgaddam,mkasula,bsamantu,ssudala,taylorad,alvarkey]@iu.edu

Indiana University-Purdue University, Indianapolis, USA

Abstract. This project is a comparative analysis and predicative model for the reporting of depression in adults. This study will investigate any discrepancies between self-reporting and healthcare professional reporting. The findings will be used to predict specific groups who are more likely to experience depression and be undiagnosed with depression.
Keywords: Depression; Reporting; NHANES database; NAMCS database; Predictive model

1 Project Scope

1.1 Introduction

People with mental health disorders are at an increased risk of social, educational, racial, and physical trouble (World Health Organization, 2022a). The incidence of depression globally in 2019 is 280 million and the World Health Organization reports an estimated 28% increase in 2020 due to the Covid pandemic (2022b). The American Academy of Pediatrics, American Academy of Child and Adolescent Psychiatry, and Children's Hospital Association declared a national mental health emergency for children and adolescents in the United States (American Academy of Pediatrics, 2021).

Despite the rise in depression and the effects of the disorder across all age groups, there remains a personal and medical stigma to a depression diagnosis (Knaak et al., 2017). Even mental healthcare providers have demonstrated stigma toward patients with mental health disorders (Hansson et al., 2013). Stigma, whether self-imposed or experienced, particularly affects the disclosure of mental health issues. Garcia et al. (2022) demonstrated that standardized screening increased the diagnosis of depression in disparaged groups. Appropriate screening in an emotionally safe healthcare setting is a critical first step to addressing the global issue of depression

The focus of this project is the screening and diagnosis of depression in adults. Due to access limitations to depression data, minors are excluded. The object of this comparison is to analyze the reporting behaviors of the adult population and diagnosis behaviors of healthcare providers regarding depression and then create a model for predicting depression reporting.

1.2 Aims

To evaluate any significant mean difference between self-reported and provider-reported cases of depression in adults in the U.S.
To identify and correlate the main factors impacting the reporting of depression by self-reporters and providers in adults in the U.S. and identify the groups that report depression the least.
To develop a multi-variant model that predicts the diagnosis of depression cases from the features of age, gender, income, co-morbidity, medical insurance, education level, and race.

1.3 Purpose

This study will investigate any discrepancies between self-reporting and healthcare professional reporting of depression and use the findings to predict specific groups, for example, a specific age range, who are more likely to experience depression and be undiagnosed with depression. The results can be used to develop screening

standards for healthcare professionals to increase the diagnosis of depression in the groups that are currently underdiagnosed.

1.4 Hypotheses:

Null Hypothesis. There will be no increase in the reporting of depression by adults as compared to the reporting of major depressive disorder by healthcare providers in medical records in the US.

Alternate Hypothesis. There will be an increase in the reporting of depression by adults as compared to the reporting of major depressive disorder by healthcare providers in medical records in the US.

2 Methodology

This project is a comparative analysis between adults and healthcare providers. The project will use two datasets to compare depression screening scores from the NHANES dataset with depression diagnosis codes from the NAMCS data. Once the data is compared, an SVM model will be developed to predict the factors impacting the reporting of depression by 1) adults and 2) healthcare providers.

2.1 Data Collection

Data from two existing publicly available datasets will be utilized for this project. Both datasets consist of secondary data collected by the National Center for Health Statistics (NCHS) and contain data from surveys and physical examinations of people in the United States (U.S.) from 2017 to 2018 (National Center for Health Statistics, 2020).

NHANES. The NCHS routinely conducts the National Health and Nutrition Examination Survey (NHANES) to gather information on the health of the general U.S. population. Nutrition, disease, health behaviors and demographic information are among the data collected from participants through a “mobile examination center” (National Center for Health Statistics, 2022b).

NAMCS. The NCHS also routinely conducts the National Ambulatory Medical Care Survey to gather information on the direct patient care received in ambulatory care centers. Examples of available data from this dataset include patient diagnosis codes, insurance coverage, provider types and practice descriptors (National Center for Health Statistics, 2018).

2.2 Data Description

This project will focus on the data within the NHANES and the NAMCS that is anticipated to correlate with depression or depression reporting (Table 1).

From the NHANES data, 11 attributes have been considered including the patient's age, patient gender, insurance type, patient race, education level, education level of adults, monthly family income, depression scale, effects of difficulties, feelings of depression, degree of depression, and comorbid conditions (See Table 1).

From the NAMCS data, the team considered 10 attributes, which includes patient age, patient gender, insurance type, patient race, provider type, specialty type, practice type, depression, chronic disease, depression screening, and mental health counseling referral (See Table 1).

Among the two data sets, the common attributes considered are patient age, patient gender, patient race, and insurance type. The team will contrast these shared characteristics across the two sets of data and look for variations in how depression is reported. With the attributes selected from the datasets, a model will be built (Support Vector Model) for future predictions about the risk of depression and depression underreporting.

Table 1. Project Variable Set (NHANES & NAMCS)[1]

VARIABLE	NHANES	NAMCS	DEPENDENT/INDEPENDENT
Patient age	Continuous	Continuous	Independent
Patient gender	Nominal	Nominal	Independent
Insurance type	Nominal	Nominal	Independent
Patient race	Nominal	Nominal	Independent
Education level	Ordinal	---	Independent
Education level-Adults 20+	Ordinal	---	Independent
Monthly family income	Continuous	---	Dependent
Depression scale (PHQ-9)[2]	Ordinal	---	Dependent
Effects of difficulties	Ordinal	---	Dependent
Feelings of depression	Ordinal	---	Dependent
How depressed?	Ordinal	---	Dependent
Comorbid conditions	Nominal	---	Dependent
Depression chronic disease	---	Nominal	Dependent
Diagnosis codes (including depression)	---	Nominal	Dependent
Provider type	---	Nominal	Independent
Specialty type	---	Nominal	Independent
Practice type	---	Nominal	Independent
Depression screen	---	Nominal	Dependent
Mental health counseling referral	---	Nominal	Dependent

[1](National Center for Health Statistics, 2018, 2022b)

[2]The Patient Health Questionnaire (PHQ-9) is a validated depression screening tool that can be self-administered. The survey consists of 9 questions and participants respond to each question on a scale of 0 (not at all) to 3 (nearly every day). A total score of 10 or greater is diagnostic of depression and is consistent with a DSM-IV depression diagnosis (Kroenke et al., 2001).

2.3 Data Storage and Extraction

Data for the various instruments in the NHANES are available on the NCHS website as data export files (xpt) along with codebooks for interpretation. The NAMCS data is also available on the NAMCS website but in a text file (txt). Using a Python panda function, the xpt and txt files will be converted to csv files and imported into a common team SQL database for storage. Python will be used to extract the variable data described in Table 1 from SQL. The team will use Github to store a team Jupyter Notebook file for Python coding.

2.4 Data Cleaning

Removing duplicates and irrelevant data. Both surveys include a broad amount of health data. This project will not analyze the full datasets. The data will be reduced to only data relevant to the project aims. Data for participants under the age of 18 will be removed. The NHANES dataset restricts the release of depression scores for minors and their scores are not included in the datasets. The project will also not analyze medications from either set. The team chose to not include this variable due to the complexity of the number of reported medications and difficulty tracing initiation and discontinuation dates.

Normalizing depression data. The depression data will be normalized to a single depression classification. For example, if one participant performs the depression screening survey and reports a chronic condition of depression, that participant will be counted once for reporting depression. If a provider reports a patient with a depression diagnosis code and a chronic condition of depression, that patient will be counted once for depression. Any indication of depression will be classified on a binary (yes/no) scale for standardization across the datasets.

Handling Missing Data. Per the NHCS (National Center for Health Statistics, 2022a) recommendation,

NHANES data will be evaluated for missing values, and no changes will be made if less than 10% are missing. Multiple imputation method will be used if more than 10% are missing. Because NAMCS data is already imputed, no adjustments are required.

2.5 Data Analysis

Descriptive Statistics.

Normality Test using Python. Measures of central tendency with variance and range will be calculated to assess distribution for each variable. Skewness and Kurtosis will be measured to assess for outliers or the need to normalize a variable.

Hypothesis testing. A two-sample t-test for groups will assess the significance of the differences between the means for self-reporting and provider-reporting. For the null hypothesis, a $p < 0.5$ will indicate rejection of the null and a significant difference between the groups.

Machine Learning and Model Testing.

Binary classification model. A Support Vector Machine (SVM) (Patel et al., 2016) will be used as a supervised machine learning model to predict between the classification of depression diagnosis or no depression diagnosis. Scikit-Learn in Python will be used to build the model as well as test the model.

*Performance Analysis***.** Python scikit-learn and a Precision-Recall (PR) Curve (Boyd et al., 2013) will be used to evaluate the model performance. The PR Curve will produce precision and recall values for probabilities at set thresholds. The PR area under the curve (AUC) will be the metric for evaluating the model's performance.

3 Deliverables and Data Visualization

3.1 Planned Visualizations using Python, Seaborn, Matplotlib and Tableau

1) Pie charts
2) Stack bar
3) Normality Test Results using Python Seaborn and Matplotlib
    a) Seaborn pairplots
    b) Histogram
4) Heatmap
5) Box plots
6) Line graph
7) Scatterplot

Stacked Bar. A stacked bar will display data between the two datasets and across groups.

Pie chart. A pie chart will display the percentages of depression reported in separate groups and in between the two main groups.

Seaborn Pairplots. The pairplots will allow the data to be visualized by pairs of features.

Histogram. A histogram will display the distribution of the depression means.

Heatmap. A heatmap will be used to view and display correlations between each variable in the datasets.

Box plots. Results of hypothesis testing will be displayed with box plots to demonstrate the means, standard deviation, and the significance of each hypothesis. Box plots will also demonstrate the descriptive statistic results for each variable (means and standard deviations) in the two datasets.

Line graph. The PR values for the PR Curve and the PR AUC will be displayed on a line graph to demonstrate the performance of the SVM model.

Scatterplot. The output from the SVM model will be displayed in a scatterplot.

4 Results

The expected outcome is that the rates of depression recorded by the public and the rates of depression reported by healthcare providers will differ. In some groups, the public may be reporting more cases of depression than healthcare providers. When compared to the depression reported by the healthcare provider, there may be a lower rate of depression recorded by the public in other groups. The analysis will provide information for the specific groups most need of screening for depression and for those providers who most need training in screening.

5 Team members Responsibility

	Team Members						
Tasks	April Taylor	Kruthika Gaddam	Bala Samantula	Sri Harsha Sudalagunta	Mohith Surya Kiran Kasula	Kiran Mai Jaiswal Charpuria	Alan Varkey
Project Management	■						■
Background/Research	■						
Proposal Development	■			■	■	■	
Editing/Proofreading	■		■			■	■
Data Collection	■	■		■	■		■
Data Analysis	■	■	■		■		
Hypothesis Testing		■		■			
Project Presentation	■	■	■	■	■	■	■
Report Development	■		■			■	
Data Cleaning		■			■	■	
Model Development and Testing	■		■				■
Data Visualization			■	■	■		■

6 Timeline

Date	Tasks
09/12/22 - 09/18/22	Data Collection and Project Pre-Draft Proposal
09/20/22 - 10/02/22	Project Draft Proposal
10/11/22 - 10/24/22	Project Final Proposal
10/25/22 - 11/04/22	Data Extraction and Data Cleaning
11/05/22 - 11/11/22	Exploratory Data Analysis
11/12/22 - 11/14/22	Hypothesis Testing
11/14/22 - 11/20/22	Risk Prediction Model
11/21/22 - 11/23/22	Forecasting
11/24/22 - 11/26/22	Data Visualization
11/27/22 - 12/01/22	Project Presentation
12/03/22 - 12/11/22	Project Report

7 References

American Academy of Pediatrics, American Academy of Child, and Adolescent Psychiatry, and Children's Hospital Association,. (2021, 10/19/2021). *AAP-AACAP-CHA Declaration of a National Emergency in Child and Adolescent Mental Health*. Retrieved 09/30/2022 from https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/

Boyd, K., Eng, K. H., & Page, C. D. (2013, 2013//). Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg.

Garcia, M. E., Hinton, L., Neuhaus, J., Feldman, M., Livaudais-Toman, J., & Karliner, L. S. (2022). Equitability of Depression Screening After Implementation of General Adult Screening in Primary Care. *JAMA Network Open*, *5*(8), e2227658-e2227658. https://doi.org/10.1001/jamanetworkopen.2022.27658

Hansson, L., Jormfeldt, H., Svedberg, P., & Svensson, B. (2013). Mental health professionals' attitudes towards people with mental illness: do they differ from attitudes held by people with mental illness? *Int J Soc Psychiatry*, *59*(1), 48-54. https://doi.org/10.1177/0020764011423176

Knaak, S., Mantler, E., & Szeto, A. (2017). Mental illness-related stigma in healthcare. *Healthcare Management Forum*, *30*(2), 111-116. https://doi.org/10.1177/0840470416679413

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

National Center for Health Statistics. (2018). *National Ambulatory Medical Care Survey*. https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NAMCS

National Center for Health Statistics. (2020). *Summary of Current Surveys and Data Collection Systems* [factsheet]. file:///C:/Users/taylo_pa2sjpo/OneDrive/Documents/I501%20Group%20Project/factsheet-summary-current-surveys.pdf

National Center for Health Statistics. (2022a). *Missing data in NHANES* Retrieved 10/24 from https://wwwn.cdc.gov/nchs/nhanes/tutorials/module1.aspx

National Center for Health Statistics. (2022b). *National Health and Nutrition Examination Survey*. National Center for Health Statistics. Retrieved 10/3 from https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2017

Patel, M. J., Khalaf, A., & Aizenstein, H. J. (2016). Studying depression using imaging and machine learning methods. *NeuroImage: Clinical*, *10*, 115-123. https://doi.org/https://doi.org/10.1016/j.nicl.2015.11.003

World Health Organization. (2022a, 11/17/2021). *Adolescent mental health*. Retrieved 10/3 from https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

World Health Organization. (2022b). *Mental Health and COVID-19: Early evidence of the pandemic's impact* (WHO/2019-nCoV/Sci_Brief/Mental_health/2022.). https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1