**FORECASTING INPATIENT MENTAL HEALTH SERVICE UTILIZATION RATES: A GOOGLE TRENDS-BASED APPROACH FOR STATE-LEVEL FORECASTING**

# **Basic**

Lead: Collins Njagi  
Recorder: Jessica Riedy  
Spokesperson: Rospide Eddy

# **Background**

This project aims to answer the following question: Can the overall popularity of mental health-related Google search trends forecast inpatient hospitalization utilization in states, after accounting for patient-level demographic data?

Mental health service utilization is a critical concern in the United States, with varying demand across states (Hogan, 2003). Accurate forecasting of its utilization is essential for healthcare planners and policy makers to allocate resources effectively (Wang et al., 2005). Traditional forecasting methods rely on historical utilization data from sources like Substance Abuse and Mental Health Service Administration (SAMHSA). However, these methods may not capture real-time forecasting, emerging trends or sudden changes in demand (Kessler et al., 2005).

This study aims to explore the potential of Google Trends data in forecasting government-run mental health service utilization at the state level. Google Trends data offers a creative approach to forecasting mental health service utilization, providing real-time data and timely insights into mental health concerns (Nuti et al., 2014; Yang et al., 2016). By investigating the relationship between Google Trends data and mental health service utilization, this study can help inform policy decisions, allocation of resources, program development, and intervention strategies, ultimately improving mental health outcomes.

While use of Google Trends data to forecast mental health service utilization is an area that has been explored before, investigating mental health utilization at the state level is still uncharted waters. In that regard, this research is somewhat novel because state level analysis can provide more nuanced insights into the specific needs and trends of each state.  
The patient population in this study consists of internet-active individuals in the U.S. who seek mental health-related information online. These individuals regularly use search engines, particularly Google, to explore mental health resources, including symptoms, treatment options, and service providers. As the study focuses on state-level trends, the population is stratified geographically based on online search behavior within each state. This group represents a subset of the U.S. population with internet access and a level of digital literacy that enables them to seek health-related information online. By analyzing their search patterns, this study aims to gain insights into mental health service utilization and improve resource allocation at the state level.

# **Hypothesis**

There is a significant positive correlation between the overall popularity of mental health-related Google search Trends and inpatient hospitalization utilization in the United States, even after accounting for patient-level demographics. This is because increased online searches for mental health information reflect growing public awareness and concern about mental health issues which in turn drives demand for mental health services, including inpatient hospitalization.

# **Forecasting**

Using Google Trends data, we forecast that states with higher overall popularity of mental health-related Google search trends for mental health-related terms (e.g., depression, ADHD) will experience higher rates of inpatient hospitalization utilization even after controlling for patient-level demographic data for the upcoming year as compared to the previous year.

# **Stakeholders**

1. State health departments: Accurate health mental service utilization forecasting will inform resource allocation and planning.

2. State policymakers and legislators: An accurate mental health service utilization forecast will inform policy and program developments.

3. Mental health service providers: Hospitals and private practices will be in a position to forecast future demand for their services.

# **Data & Methods**

The Google Trends and SAMHSA datasets are a good choice for this research question because they provide complementary information on mental health service utilization. The curated [Google Trends data](https://docs.google.com/spreadsheets/d/1yxxDpVBgw2bQId5XmdslYMYOFBgVqEM0/edit?usp=sharing&ouid=101501188568290214846&rtpof=true&sd=true) offers insights into online search behavior related to mental health topics, which can serve as an indicator for mental health concerns and service utilization. On the other hand, [the Mental Health Treatment Services Dataset](https://www.samhsa.gov/data/data-we-collect/mh-cld-mental-health-client-level-data/datafiles) is a Public-Use File (PUF) released by the Substance Abuse and Mental Health Service Administration (SAMHSA) that includes the Mental Health Client Level Data (MH-CLD) dataset and provides actual utilization data on mental health services at the state level. By combining these two datasets, we can examine the relationship between online search behavior and service utilization, allowing us to assess the predictive power of Google Trends data.

To merge the datasets, we will use state and year as our common identifier. The curated Google Trends data will be read into R using the `*readxl*` library and then merged with the SAMHSA dataset using the `*inner\_join*` function from the `*dplyr*`library. Basic cleaning steps will include handling missing values, removing duplicates, and ensuring consistency. Our analysis will be conducted in R for manipulation and modeling. Our response feature will be inpatient utilization rates (excluding the private inpatient since our focus is government-run institutions) from the curated Google Trends dataset and our predictor features will be all the rest of the variables from our merged datasets.

**Analysis Plan**

## **A. Data Preparation**

This will be a crucial step in ensuring the quality and reliability of the analysis.

### i. Data Import and Inspection: Google Trends and SAMHSA datasets will be imported into RStudio for analysis. These datasets will be inspected for missing values, obvious errors, inconsistencies, and variables related to mental health service utilization.

ii. Handling missing values: For Google Trends, missing values will be handled using the spline interpolation method, while SAMHSA will employ predictive mean matching.

iii. Exploratory data analysis: To understand the distribution of the data, identify possible outliers, and get a feel of relationships between variables, the following methods will be used;

· Summary statistics: We will calculate the mean, median, and standard deviation.

· Visualization using plots to understand distribution.

· Identify outliers.

iv. Feature Engineering: This will involve creating new features from the existing features to improve accuracy and reliability. For Google Trends, this will involve search terms closely related to mental health, focusing on specific states and years. Examples include depression, ADHD, and Down syndrome.

For SAMHSA, we will create new variables, for example, funding allocation for mental health services by state and year.

v. Data Aggregation and Merging: We intend to merge curated Google Trends and SAMHSA datasets using State and year variables to create a single dataset for analysis.

1. Filename:
2. File Descriptions:
3. Please, see the [data dictionary](https://docs.google.com/spreadsheets/d/1SXHXBpQRFhJkHSFj3jGkuEjQF0Lj5YrZF9oD_WzKXBY/edit?usp=sharing).

## **B. Model Development**

### **Model 1: Machine Learning with Random Forest**

· Model specification: We intend to use Random Forest regression to model the relationship between Google Trends features and SAMHSA utilization rates.

· Hyperparameter Tuning: We will try to optimize the model hyperparameters using a random search method.

· Feature importance: We will use an inbuilt Random Forest feature importance tool to assess the most important features.

· Model Evaluation: We will use Mean Absolute Error (MAE), R-squared, and Mean Squared Error (MSE) to evaluate the performance of the model.

### **Model 2: Linear Regression with Elastic Net Regularization (Elastic Net Regression)**

o Split data into train dataset (70%) and test dataset (30%)

o Create an Elastic net regression model: We intend to use `*glmnet*`library in R

o Tune hyperparameters: We will use cross-validation to tune the hyperparameters

o Train the model on the train dataset

o Model evaluation: We intend to evaluate the model performance on our test dataset using MAE or MSE metrics

## **C. Model comparison and selection**

We will use MAE, MSE, and R-squared as our comparison metrics. The model with the best performance will be selected.

## **D. Forecasting and Validation**

The selected model will be put into use to forecast mental health service utilization rates for the upcoming year. It will then be validated using MAE, MSE, and R-Squared metrics.

# **How will you know if your question is answered?**

* High forecasting accuracy of the model in forecasting mental health service utilization rate using Google Trends data.
* Consistent patterns in relationships between Google Trends data and mental health service utilization rates across different states within the same period.

# **How will you know if your hypothesis is supported?**

* There is a high correlation between Google Trends data and mental health service utilization (r)^2 > 0.7.
* Low Root Mean Square Error (RMSE < 10) great performance,
* 10 <=RMSE < 20 will be acceptable model
* RMSE >= 20 will be a poor model

# **References**

Hogan, M. F. (2003). The president’s New Freedom Commission: Recommendations to transform mental health care in America. Psychiatric Services, 54 (11), 1467-1574.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62 (6), 593-602.

Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., & Murugiah, K. (2014). The use of Google Trends in health care research: A systematic review. PLOS ONE, 9(10), e109583.

Wang, P. S., Lane, M., Olfson, M., Pincus, H. A., Wells, K. B., & Kessler, R. C. (2005). Twelve-month use of mental health services in the United States: Results from the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6), 629-640.

Yang, A. C., Huang, N. E., Peng, C. K., & Tsai, S. J. (2016). Do seasons have an impact on the incidence of depression? The use of Google Trends in a population-based study. Depression and Anxiety, 33(7), 578-585.

# **Appendix**