NHANES Reimagined

A complex data-driven approach to analyzing public health trends

Aaron Adams (aadams61), Nafisa Barlaskar (nbarlaskar3), Kevin Caron (kcaron7), Kevin Chen (kchen604), John Osborn (josborn9), Sidharth Parwani (sparwani6)

Introduction & Survey

The National Health and Nutrition Examination Survey (NHANES) is a cross-sectional program of studies administered by the U.S. Centers for Disease Control and Prevention (CDC) designed to assess the heath and nutritional status of the non-institutionalized population in the United States beginning in the 1960's [1]. While the study was originally conducted at intermittent time intervals, in 1999 NHANES became a continuous survey allowing for the collection of a nationally representative sample of approximately 10,000 different individuals assessed from approximately 15 different locations around the U.S. with several thousand health variables examined per person every 2-year period [2]. NHANES is considered the cornerstone for national nutrition monitoring and is regularly used to guide nutrition and health policy throughout the United States [3, 4].

NHANES data is frequently applied to public health research and analyzed in scientific publications examining a wide range of topics [5] such as U.S. secondhand smoke exposure [6], asthma prevalence [7], blood pressure and risk of stroke [8, 9], obesity [10], as well as socioeconomic health disparities based upon race [11], sex [12] and combinations of many other demographic variables [13]. The objective of our project is to make NHANES data analysis and visualizations more accessible and easier to interpret for public health workers, policy makers, and members of the public. Currently, NHANES data sets representing individual variables or groups of variables for each 2-year period are released in SAS export files (.XPT) located on the NHANES website. The current structure requires users to individually download the files before viewing, analyzing, and visualizing the data with the SAS statistical software suite or another application requiring conversion of the files into different formats [14]. Our team created a dynamic program using Python web scraping tools and SQLite that iterates through each .XPT file on the NHANES website and creates a complete database containing all of the continuous NHANES data.

Visualizations form an important part of public health informatics communications [15]. Visualizing data facilitates discussion, aids understanding, makes patterns apparent, promotes analysis, and fosters recall [16, 17]. Currently, publicly available visualizations of NHANES data are limited to line graphs of an extremely small subset of variables related to total cholesterol, hypertension, and obesity with the ability to breakdown information by race, age group, and gender [18]. Furthermore, attempts to visualize NHANES data in the literature require paid access to journals and programs focused on very specific types of analysis such as multi-dimensional scatter plots of dietary data [19]. The current dashboard maintained by CDC is limited to representing data in two dimensions and the dashboard does not represent all of the available data, does not allow for aggregation of data points, and does not give the user the ability to drill down into the data for further investigation. It is therefore our assessment, that in its current form, NHANES data lacks a public, freely accessible interface that employs modern data mining methods to glean meaningful patterns and trends [20]. Therefore, in addition to creating a comprehensive database, we employ modern data mining techniques to prepare the data for advanced analysis and publish an interactive dashboard for visualizing thousands of NHANES results related to demographic, medical examination, laboratory, and questionnaire variables across survey periods[21, 22].

Problem Definition

This project seeks to address three main problems of accessibility and functionality with the existing NHANES dashboard, with the end goal of providing a comprehensive, easy-to-use, and insightful visualization tool available for anyone looking to use NHANES data for analysis or exploration.

First, the data exists on the NHANES website as SAS export files, fragmented by survey period, component (lab results, demographic information, questionnaire responses), and by subcomponents. In their given state, the files require SAS or the use of some program/basic code to convert files into a more user-friendly file type. For example, at present, if someone wanted to evaluate total cholesterol levels in the US population since 1999, they would need to download, convert, and merge over 12 data files. Streamlining this process and increasing accessibility is a top priority for this project.

Second, the official dashboard currently supports visualizing three topics (cholesterol, hypertension, and obesity) with filtering by only two demographic components (age and race). This is a major disservice to a data set that has the capability of providing rich and complex insight into nationwide public health trends.

Third, current visualization tools for accessing NHANES data provide limited data analysis, making it difficult for users to derive meaning. NHANES is a complex, multistage probability study and, with the proper application of appropriate statistical tools, can be used to estimate health metrics that are representative of the US population. The current visualizations simply summarize NHANES participant data without the application of statistical techniques needed to reveal the true impact of the data on national health. Additionally, the current dashboard fails to provide insight into the future of United States public health. We intend to build an application that addresses these needs and opens the door to further analysis and development.

Proposed Method

We tackle these deficiencies by implementing the following innovations: 1. Using Python scripts to download, sort, clean, and merge the data into a more accessible and complete data set 2. Host the data along with an interactive dashboard 3. Apply sample weights so the data presented is statistically representative of the US population 4. Add regression and forecasting/smoothing techniques to provide insight into future trends 5. Predict the occurrence of medical conditions based on the test results and user input. The first two innovations aim to tackle the accessibility problem outlined in the problem defined above. Compiling and preprocessing the data will remove this burden from end-users wanting to take a deeper dive into the data. We have identified multiple useful software tools that have many built-in interactive capabilities. We aim to fully utilize this software's mouse-over, auto-coloring, slider, and play button functionality to aid in providing insight via user interaction. By taking advantage of the software's built-in features, we can spend more time setting up meaningful dashboard components and deriving meaning from the wealth of data captured in the NHANES data set.

Unlike the commonly used simple random sample approach, NHANES implements a complex, multistage probability sampling design to select a sample that is statistically representative of the civilian non-institutionalized resident population of the United States. The proper application of this approach is not trivial, requiring the selection of the appropriate weights for each variable, as well as utilization of the masked strata and primary sampling units (PSUs) associated with these variables. This information is then applied to the raw sample data (in the case of NHANES each survey period sample size is approximately 10,000 people) to make it statistically representative of the entire US population [23] (approx. 332,000,000) [24]. This process corrects for under/over representation of sub-populations in the actual sample while preserving accurate analysis of minority populations. The raw data presented on the website contains the information needed to apply the calculation for sample weighting, but it is not applied by default. Our approach will follow the NHANES Analytic Guidelines using edited versions of the samplics python package to conduct this analysis for creating weighted histograms, arithmetic means, geometric means, and 95% confidence intervals [2], [25]. Insight into the past and present is useful information, but being able to predict/forecast future public health outcomes provides yet another dimension of analysis from this data set. The current official visualization does not allow for prediction or forecasting, two useful analytic tools. We hope that by implementing short-term forecasting methods such as exponential smoothing or ARIMA, we can provide a tool for public health trend prediction.

At a high level, we aim to provide an interactive dashboard (discussed in the following two sections) with multiple features that each facilitate a different type of analysis. For example, a dashboard component that allows the end-user to select a laboratory variable (which would already be properly weighted) and to visualize how its distribution changes over time with a play button/slider. This would provide a deeper understanding of how a population is changing in a way that simple summary statistics cannot capture. Another section will be focused on providing predicted trends using forecasting models. Providing a component to run linear regression models between two continuous variables would also provide useful insight into how variables are related to one another. Also, performing machine learning algorithms on these data might help in predicting health conditions based on the test results in the NHANES dataset. These are a few examples of useful, non-trivial analyses that we expect to implement.

Experiments & Evaluation

How to Organize the Database?

The first stage of experimentation and evaluation was making sure we could get the data needed for building the dashboard into a useful form. This required looping over the downloadable files on the NHANES website and downloading and storing the data in a data frame. Here we are testing three approaches to storing the data.

We tested three forms of data set structure. The first form was a long data table, where every unique response/lab result would be stored as a row in the table. The second form was storing the information in a collection of wide data frames, each representing a survey period, where the rows are all unique sequence numbers from that survey period, and the columns are all potential demographic/questionnaire/laboratory variables. The third form was created by taking a subset of columns that are present across all survey periods, so we could stack the wide data frames into a single table. This would result in the loss of some variables. However, the approximately 100 variables that are present across all survey periods imply a level of importance and greatly expand upon the existing 3 variables in the current dashboard. Ultimately, the third form was chosen as it was much faster to query than form 1, and it was much easier to incorporate one database into a visualization rather than dynamically querying from 10 tables.

Best Visualization Framework to Use?

In the project proposal, we identified Tableau as our front-runner for building the visualization. After testing, we were able to load data into a Tableau dashboard, but performing sample weighting, regression, forecasting, and other non-trivial analyses within Tableau had many roadblocks. As such, we found a suitable replacement using interactive Jupyter Notebook widgets. The widgets have many of the features that drew us towards Tableau in the first place (drop-downs and slider bars), making it a great alternative. This approach incorporated the working environment from many of our homework assignments, and facilitated more direct developer-friendly control for creating functions and models with python and its many useful libraries. Pairing widgets with graphical libraries like matplotlib allowed for more control over the final dashboard components and flexibility with design.

Extensive experimentation has been done to determine how best to organize our visuals in a way that tells a clear and compelling story for the end user. Testing for the legibility of graphics, interactivity between widgets, and general ease of use throughout the dashboard was the primary focus of our User Experience Testing. Additionally, more experimenting was done towards refining and modeling the data, specifically with regards to which forecasting techniques are used (smoothing vs ARIMA) as well as which predictive models are best suited for this data set (Support Vector Machine, regression, clustering).

Which Models to Use for Predicting Survey Values?

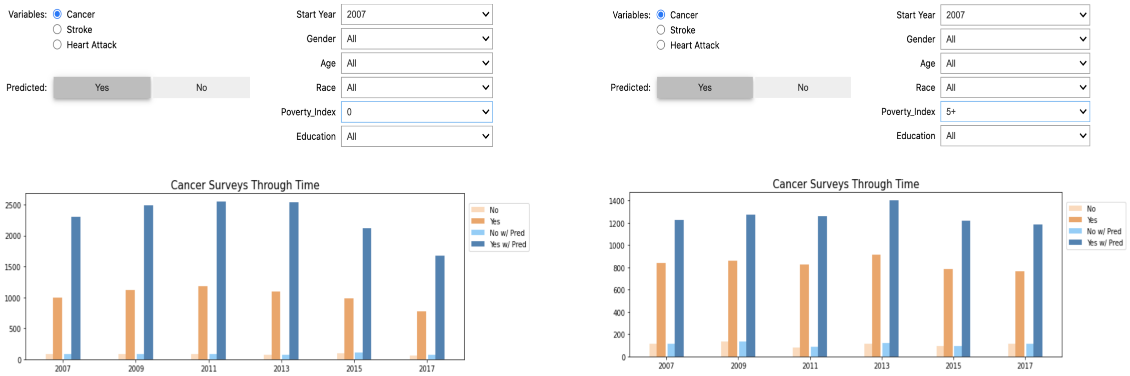
We implemented a Holt-Winters forecasting model in order to predict weighted population values 1 to 3 survey periods in the future. A Holt-Winters model was chosen as it allows the user to adjust the level (tendency to the average value in the series) and the trend (recent increasing or decreasing in the series). Seasonality was not useful for the short time series of 10 data points. Users can select a domain (race or gender) to observe potential public health disparities between sub-populations in the past, present, and future. Furthermore, we built a scatter plot widget and a simple linear regression model widget that allows users to visualize and quantify the relationship between variables in the NHANES dataset.

How to Reduce Number of Training Variables?

When deciding on variables to include in our revamped visualization, it was noticed that free response data was missing (or reported as ”Don’t Know”) in the range of 40-50% depending on the variable, namely those concerning self-reporting prior diagnoses of Stroke, Heart Attack, or Cancer. This showed itself as an opportunity for the team to expand on the project goal of informing policymakers and increasing public knowledge with more accurate health information. Specifically, we sought to predict the answers to these missing free-response questions to reveal a more complete representation of cardiovascular disease and cancer in the US. The resulting visualization can be seen below.

One main hurdle encountered during the creation of this visual was the number of possible variables to train the prediction model. As mentioned, there were more than one hundred possible training variables, increasing the risk of overfitting the training data. Intent on avoiding this risk, principal component analysis was performed on the data set to reduce the number of variables used in the model. Variables found to be consistently contributing less than 10% to any of the six principal components found were removed until less than twenty training variables remained. The remaining variables still captured 90% of the variance in the data and enabled the prediction accuracy of our random forest model to achieve approximately 92% testing accuracy for 2 out of 3 variables which is on par with fitting with all variables.

The results allow users to see how many missing responses resulted in ”Yes” predictions from our model, suggesting a higher rate of cardiovascular disease and cancer among those who did not respond than the raw survey data would show initially, even after the removal of the empty responses. Additionally, it’s possible to add another dimension to this conclusion by slicing the data by poverty levels. When doing so, we see that this gap between predicted and non-predicted ”Yes” answers is substantially higher for lower-income individuals(Figure 1a) than for those with



Variables: ● Cancer ○ Stroke ○ Heart Attack

Predicted: Yes No

Start Year: 2007

Gender: All

Age: All

Race: All

Poverty_Index: 0

Education: All

(a) Low Wealth Populations

Variables: ● Cancer ○ Stroke ○ Heart Attack

Predicted: Yes No

Start Year: 2007

Gender: All

Age: All

Race: All

Poverty_Index: 5+

Education: All

(b) High Wealth Populations

Figure 1: Comparing Cancer Rates Between High and Low Wealth Populations

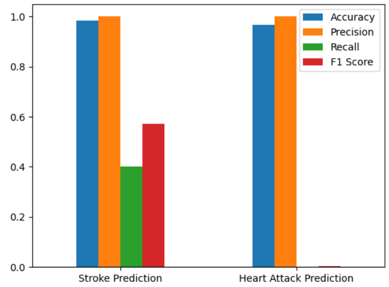
higher income(Figure 1b). This striking conclusion is just one of many that are now possible to the public with the myriad dimensions available to drill down on the data with.

How can we visualize population distributions of health metrics over time and between demographic groups?

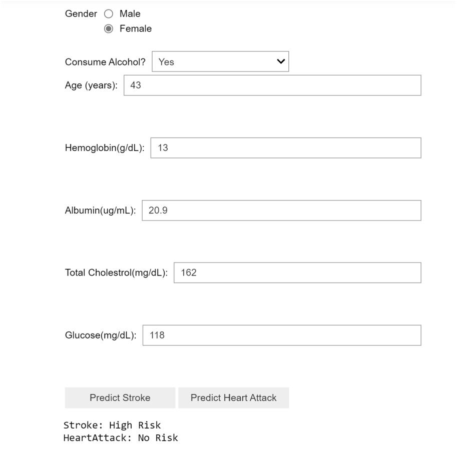
One feature we designed was utilizing the population weight conversion to determine the population frequency distribution of a specific variable during a given survey year. After designing this, we implemented a way to export the year-by-year plots to a .gif file format to display the distribution as in a time series. This allows the user to see trends using frames as different survey year plots rather than quantified data outputs like moving averages and standard deviations. This model makes it easier to tell a story by using images rather than using raw data since explicit data can sometimes be difficult to comprehend by the public. This model also helps visualize and better understand more complex models. For example, the LBXCOT measures the Serum Cotinine concentration of the surveyee. Serum Cotinine is a compound formed when nicotine is metabolized by the body and is very frequently observed as a bimodal distribution across the population, with higher concentrations associated with smokers and lower concentrations associated with nonsmokers exposed to secondhand smoke. This is very obvious to see using the graphic, but would be difficult to discern a trend if using raw data. Additionally, we implemented a feature allowing users to compare these distributions in histograms by population subgroup based on race and sex.

How do we use the data to predict conditions like Heart Attack, Stroke, etc.?

We analyzed the NHANES data to find trends and predict if a patient is at risk for heart attack or stroke by performing a classification of patients who have been diagnosed with a stroke or heart attack at least once by using machine learning. We researched the variables that impact the likelihood of heart attack and stroke such as albumin [26], cholesterol[27], age and gender[28], alcohol[29], glucose [30] etc. Based on this research, we identified and extracted the variables like RIDAGEYR (age), RIAGENDR (gender), PHQ030 (Consuming Alcohol), LBXHGB (Hemoglobin), URXUMA (Albumin), LBXTC (Total Cholesterol), and LBXSGL (Glucose) from the dataset. The dataset had sparse values for several other columns such as triglycerides etc and those were discarded to improve accuracy. The MCQ questions MCQ160F and MCQ160E, in which the participants have answered with either Yes (1), No (2), Refused (7), Don't Know (9), etc., were used as labels to



(a) Stroke and Heart Attack prediction metrics



Gender ○ Male
● Female

Consume Alcohol? Yes

Age (years): 43

Hemoglobin(g/dL): 13

Albumin(ug/mL): 20.9

Total Cholestrol(mg/dL): 162

Glucose(mg/dL): 118

```
Stroke: High Risk
HeartAttack: No Risk
```

(b) UI to predict based on user inputs

Figure 2: Comparing Cancer Rates Between High and Low Wealth Populations

predict stroke and heart attack in patients respectively. The cleaned dataset and labels were used to build two Support Vector Machine (SVM) classifiers. The accuracy obtained with these two classifiers was 97% and 96% percent respectively.

SVM was used because it achieved higher accuracy over other classifiers and it performs better when the distribution of the data is irregular or unknown. It maximizes the distance between the margins and is able to handle high dimensional data and mitigates overfitting.

A UI tool was built using iPywidgets, which takes in the test results as input from users and it predicts their heart attack or stroke probability. The input values are run against the two classifiers as per the button clicked by the user.

Please note that the results of the prediction are not comprehensive, and require further research and analysis to come to a holistic conclusion. The correlation between these variables with stroke and heart attack occurrence has been studied extensively, yet results are still inconclusive. This experiment was an attempt to demonstrate that real datasets like the NHANES can be used for medical analysis and research by relying on modern machine-learning techniques to get better outcomes and conclusions.

Conclusions & Discussion

We have successfully demonstrated that a complex dataset like NHANES can be used to predict medical trends, obtain summary statistics of continuous variables, explore population-wide distribution of health metrics. We also used this data to build machine learning models to predict medical conditions. Through our interactive widgets we have demonstrated the power of visual analytics and machine learning in the context of developing a deeper understanding on public health trends in the United States.

Some future work for this project will include finding a way to host the interactive widgets as we recognize running a local jupyter notebook file is prohibitive of reaching a wide scale audience, as well as continual development and improvement of existing models.

All team members have contributed an equal amount amount of effort towards completion of this project.

References

[1] L. Curtin, L. Mohadjer, S. Dohrmann, J. Montaquila, D. Kruszan-Moran, L. Mirel, M. Carroll, R. Hirsch, S. Schober, and C. Johnson, "The national health and nutrition examination survey: Sample design, 1999-2006," *Vital and health statistics. Series 2, Data evaluation and methods research*, 2012.

[2] C. Johnson, R. Paulose-Ram, C. Ogden, M. Carroll, D. Kruszon-Moran, S. Dohrmann, and L. Curtin, "National health and nutrition examination survey: analytic guidelines, 1999-2010," *Vital and health statistics. Series 2, Data evaluation and methods research*, 2013.

[3] N. Ahluwalia, J. Dwyer, A. Terry, A. Moshfegh, and C. Johnson, "Update on NHANES Dietary Data: Focus on Collection, Release, Analytical Considerations, and Uses to Inform Public Policy," *Advances in Nutrition*, vol. 7, no. 1, pp. 121–134, 01 2016. [Online]. Available: https://doi.org/10.3945/an.115.009258

[4] J. A. Fain, "NHANES: Use of a Free Public Data Set," *The Diabetes Educator*, vol. 43, no. 2, mar 2017. [Online]. Available: https://doi.org/10.1177%2F0145721717698651

[5] P. Eke, G. Thornton-Evans, L. Wei, W. Borgnakke, and B. Dye, "Accuracy of NHANES periodontal examination protocols," *Journal of Dental Research*, vol. 89, no. 11, pp. 1208–1213, sep 2010. [Online]. Available: https://doi.org/10.1177%2F0022034510377793

[6] K. T. Caron, W. Zhu, J. T. Bernert, L. Wang, B. C. Blount, K. Dortch, R. E. Hunter, T. Harmon, J. R. Akins, J. Tsai, D. M. Homa, J. L. Pirkle, and C. S. Sosnoff, "Geometric mean serum cotinine concentrations confirm a continued decline in secondhand smoke exposure among u.s. nonsmokers—NHANES 2003 to 2018," *International Journal of Environmental Research and Public Health*, vol. 19, no. 10, p. 5862, may 2022. [Online]. Available: https://doi.org/10.3390%2Fijerph19105862

[7] M. K. McHugh, E. Symanski, L. A. Pompeii, and G. L. Delclos, "Prevalence of asthma among adult females and males in the united states: Results from the national health and nutrition examination survey (NHANES), 2001–2004," *Journal of Asthma*, vol. 46, no. 8, pp. 759–766, jan 2009. [Online]. Available: https://doi.org/10.1080%2F02770900903067895

[8] D. BROWN, W. GILES, and K. GREENLUND, "Blood pressure parameters and risk of fatal stroke, NHANES II mortality study," *American Journal of Hypertension*, vol. 20, no. 3, pp. 338–341, mar 2007. [Online]. Available: https://doi.org/10.1016%2Fj.amjhyper.2006.08.004

[9] P. W. Yoon, B. Bastian, R. N. Anderson, J. L. Collins, and H. W. Jaffe, "Potentially preventable deaths from the five leading causes of death – united states, 2008–2010," *Morbidity and Mortality Weekly Report*, 2014.

[10] C. A. Befort, N. Nazir, and M. G. Perri, "Prevalence of obesity among adults from rural and urban areas of the united states: Findings from NHANES (2005-2008)," *The Journal of Rural Health*, vol. 28, no. 4, pp. 392–397, may 2012. [Online]. Available: https://doi.org/10.1111%2Fj.1748-0361.2012.00411.x

[11] A. Y. Rosinger, A. I. Patel, and F. Weaks, "Examining recent trends in the racial disparity gap in tap water consumption: NHANES 2011–2018," *Public Health Nutrition*, pp. 1–7, jun 2021. [Online]. Available: https://doi.org/10.1017%2Fs1368980021002603

[12] E. B. Loucks, D. H. Rehkopf, R. C. Thurston, and I. Kawachi, “Socioeconomic disparities in metabolic syndrome differ by gender: Evidence from NHANES III,” *Annals of Epidemiology*, vol. 17, no. 1, pp. 19–26, jan 2007. [Online]. Available: https://doi.org/10.1016%2Fj.annepidem.2006.07.002

[13] X. Zhang, M. F. Cotch, A. Ryskulova, S. A. Primo, P. Nair, C.-F. Chou, L. S. Geiss, L. E. Barker, A. F. Elliott, J. E. Crews, and J. B. Saaddine, “Vision health disparities in the united states by race/ethnicity, education, and economic status: Findings from two nationally representative surveys,” *American Journal of Ophthalmology*, vol. 154, no. 6, pp. S53–S62.e1, dec 2012. [Online]. Available: https://doi.org/10.1016%2Fj.ajo.2011.08.045

[14] (NaN) Nhanes - national health and nutrition examination survey homepage. [Online]. Available: https://www.cdc.gov/nchs/nhanes/index.htm

[15] E. McGill, V. Er, T. Penney, M. Egan, M. White, P. Meier, M. Whitehead, K. Lock, R. A. de Cuevas, R. Smith, N. Savona, H. Rutter, D. Marks, F. de Vocht, S. Cummins, J. Popay, and M. Petticrew, “Evaluation of public health interventions from a complex systems perspective: A research methods review,” *Social Science &amp Medicine*, vol. 272, p. 113697, mar 2021. [Online]. Available: https://doi.org/10.1016%2Fj.socscimed.2021.113697

[16] K. Narayan and M. Nayak, “Need for interactive data visualization in public health practice: Examples from india,” *Int J Prev Med.*, 2021.

[17] S. R. Minshall, H. Monkman, A. Kushniruk, and L. Calzoni, “Towards the adoption of novel visualizations in public health,” in *Studies in Health Technology and Informatics.* IOS Press, jun 2022. [Online]. Available: https://doi.org/10.3233%2Fshti220680

[18] N. Ansai and A. Lipphardt. (NaN) Nhanes - interactive data visualizations. [Online]. Available: https://www.cdc.gov/nchs/nhanes/visualization/index.htm#dashboard

[19] S. O. Torres, H. Eicher-Miller, C. Boushey, D. Ebert, and R. Maciejewski. (2012) Applied visual analytics for exploring the national health and nutrition examination survey. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6149111

[20] Z. Xing, “Exploring disease association from the nhanes data,” *International Journal of Data Warehousing and Mining*, vol. 6, pp. 3–36, 2010.

[21] S. Kaushik, A. Choudhury, P. K. Sheron, N. Dasgupta, S. Natarajan, L. A. Pickett, and V. Dutt, “AI in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures,” *Frontiers in Big Data*, vol. 3, mar 2020. [Online]. Available: https://doi.org/10.3389%2Ffdata.2020.00004

[22] I. N. Soyiri and D. D. Reidpath, “An overview of health forecasting,” *Environmental Health and Preventive Medicine*, vol. 18, no. 1, pp. 1–9, jul 2012. [Online]. Available: https://doi.org/10.1007%2Fs12199-012-0294-6

[23] A. C. Skinner, S. N. Ravanbakht, J. A. Skelton, E. M. Perrin, and S. C. Armstrong, “Prevalence of obesity and severe obesity in US children, 1999–2016,” *Pediatrics*, vol. 141, no. 3, mar 2018. [Online]. Available: https://doi.org/10.1542%2Fpeds.2017-3459

[24] (NaN) Population clock. [Online]. Available: https://www.census.gov/popclock/

[25] M. S. Diallo, "samplics: a python package for selecting, weighting and analyzing data from complex sampling designs." *Journal of Open Source Software*, vol. 6, no. 68, p. 3376, 2021. [Online]. Available: https://doi.org/10.21105/joss.03376

[26] L. C. Y. H. Chien SC, Chen CY, "Critical appraisal of the role of serum albumin in cardiovascular disease." *Biomark Res 5*, Nov 2017. [Online]. Available: https://doi.org/10.1186/s40364-017-0111-x

[27] H. S. L. R. L. W. J. P. B. Tirschwell DL, Smith NL, "Association of cholesterol with stroke risk varies in stroke subtypes and patient subgroups." *Neurology*, Nov 2004. [Online]. Available: https://doi.org/10.1212/01.wnl.0000144282.42222.da

[28] A. Y. Y. C. C. J. H. L. Kathryn M. Rexrode, Tracy E. Madsen and E. C. Miller, "The impact of sex and gender on stroke." *Circulation Research*, Feb 2022. [Online]. Available: https://doi.org/10.1161/CIRCRESAHA.121.319915

[29] B. D. Emberson JR, "Effect of alcohol on risk of coronary heart disease and stroke: causality, bias, or a bit of both?" *Vasc Health Risk Manag*, Sep 2006. [Online]. Available: https://doi.org/10.2147%2Fvhrm.2006.2.3.239

[30] F. W. Chen R, Ovbiagele B, "Diabetes and stroke: Epidemiology, pathophysiology, pharmaceuticals and outcomes." *Am J Med Sci*, Apr 2016. [Online]. Available: https://doi.org/10.1016%2Fj.amjms.2016.01.011