**NHANES Diabetes and Demographic data 2017-2018:**

The National Health and Nutrition Examination Survey (NHANES) is designed to collect/sample data to assess the health and nutritional status of adults and children the US. The data obtained through survey which is unique in that it combines interviews, physical examinations, and administers tests of physical activity and fitness. NHANES collects person-level demographic, health, and nutrition information. It produces national estimates that are representative of the total non-institutionalized civilian U.S. population. The NHANES sample is selected using a complex, four-stage sample design.

Why 2017-2018 chosen – (Ref. 1): The coronavirus disease 2019 (COVID-19) pandemic required suspension of data collection in March 2020. As a result, the partially completed NHANES 2019–2020 cycle was not nationally representative. Therefore, the 2019–March 2020 data were combined with the data from the 2017–2018 cycle to create the nationally representative 2017–March 2020 pre-pandemic data files and was published only in 2022. Datasets for 2021-2023 are yet to be released as of Jan 15, 2024.Therefore, decided to go for the 2017 -2018 cycle dataset which is more reliable and collected without any interruption.

Why Demographic and Questionnaire data – Examination and Laboratory data are used and analyzed in several studies using Electronic Health Record (EHR) data. Inquisitive about using this easily available, tested, self-reported, accurate, and reliable data from CDC/NHANES. Not many studies are done using only this questionnaire data.

**Data sources**: from public use data files <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&Cycle=2017-2020>

<https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2017-2020>

**Data descriptions**: <https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.htm>

<https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DIQ.htm>

Analytic guidelines and Reporting: https://wwwn.cdc.gov/nchs/data/nhanes/analyticguidelines/05-06-analytic-guidelines.pdf

**Query for: (Results)**

No. of record in 2017-2018 demographic\_questionnaire file = 9254

No. of records in 2017-2018 diabetic\_questionnaire file = 8897

% of respondents who have Diabetes

% of respondents who are Pre-diabetes

Distributions of Age, Gender, Race, Ethnicity, Edn. level, Income, Pregnancy status, Language, INDHHIN2 - Annual household income, INDFMPIR - Ratio of family income to poverty – For all Diabetics

Distributions of Age, Race, Edn level, Income of patients who are Pre – DM 🡪 DM and

those who are Pre – DM 🡪 No DM.

Regression analysis for patients who developed Diabetic Retinopathy.

**Previous studies on Diabetic Retinopathy:**

1. Prevalence of Diabetic Retinopathy in the US, 2005 -2008. JAMA. 2010;304(6):649-656. doi:10.1001/jama.2010.1111 (U of Wisconsin)

Used NHANES data from 2005 -2008, 40 years or older adults, approval by human subjects’ review board, written informed consent by all participants, digital images of retina were taken for all participants and analyzed. **Multiple Logistic regressions** were used to assess association between Diabetic retinopathy Vs clinical potential risk factors.

1. Prevalence of Diabetic Retinopathy in the US in 2021.

JAMA Ophthalmol. 2023;141(8):747-754. Doi:10.1001/jamaophthalmol.2023.2289

Used NHANES data from 2005-2008, 2017 -March 2020 data, Medicare fee-for-service claims (2018), IBM MarketScan commercial insurance claims (2016), population-based studies of adult eye diseases (2001 to 2016), studies of diabetes in youth (2021 and 2023), and a previously published analysis of diabetes by county (2012). **Bayesian meta-regression** methods used.

We can produce the same kind of results like [1] study done in U of Wisconsin using the more recent, publicly available data source which only involved NHANES questionnaire data on Diabetes, Demographics, Medications for the study period. Avoiding the arduous processes of IRB approval, patient consent, additional examinations. Even Retinal fundus photography may underestimate the prevalence of Diabetic retinopathy as OCT (Optical Coherence Tomography) is better at detecting these conditions.

Also, one study on diabetic retinopathy might look a lot better if it is backed up by our study, or if our study says differently, could be eye-opening to the clinical research community about other considerations etc.,.

If you need to reinstall the xport package: *python3 -m pip install xport=2.0.2 –user*

If you need to upgrade pip: *python3 -m pip install –upgrade pip --user*

If you need to add a new script to PATH: *PATH = %PATH%;directory/executable.app*

To list what folders are in a directory: *dir*

To set that as the new current directory: *cd*

To convert a xpt file to csv: *python3 -m xport filename.xpt > filename.csv*

Using temporary access token (only available till August 22, 2024) for GitHub access, remote access denied otherwise. Remote access token: github\_pat\_11AHW5FXQ0xa6MO5kiuu5F\_n7M5CEY5ZNLeEm8dPifNJmQdlC1q1NgBQiNY9gq38IDIKDDKK4OKkwQCiZr

Sometimes the missing values in the data include inapplicable cases (such as A1C reported as missing for non-diabetics). Point is, the data has to make sense. When we ask if diabetics are all diagnosed with the same set of questions regardless of demographics, we need to make sure the information isn’t missing or unaccounted simply because they’re non-diabetics.

We need to ensure that when we look at each kind of question, we’re working with rows in the data where that question applies, and that the column information is appropriately identified in a subset. Fortunately, we have every individual’s id/sequence number.

We took the filtered table of only people with diabetes (with their demographic info included), we checked the proportion of each ethnicity/race among the diabetic patients (compared with overall surveyed proportions), we also checked proportion of diabetic patients that belonged to each ethnicity/race AND education level (compared to the proportions seen by the overall survey).

This will allow us to do a Pearson Chi-square test to see if the distribution of race for diabetics is the same as the overall surveyed, and likewise with the distribution of race and education for diabetics vs overall.

The purpose of the Chi-square tests is to see which demographic variable’s distribution has a statistically significant difference to the overall survey distribution for that demographic variable. That way we know which variables to focus on in our plots. We can also use that to reduce the complexity of any linear models involving demographics and diabetes (or pre-diabetes or diabetic retinopathy).

Example of plots to have in paper: <https://medium.com/the-researchers-guide/introduction-to-dodged-bar-plot-matplotlib-pandas-and-seaborn-visualization-guide-part-2-1-49e2fbc9ac39>

References:

1. <https://www.cdc.gov/nchs/data/series/sr_02/sr02-190.pdf>