

# ENAR 2024 DataFest Project Report

Mailman School of Public Health at Columbia University

Yi Huang, Yiying Wu

Dec 21, 2023

## **Abstract**

This study used 1999-2020 National Health and Nutrition Examination Survey (NHANES) data to investigate hypertension risk factors in the U.S. adult population. We aim to identify the potential causes or correlates of worsening BP control among US adults with hypertension over the past decade. Logistic regression analysis incorporated multi-year dataset weighting and multiple imputations for missing data. Key variables included survey year, age, race, gender, BMI, diabetes, CKD, CVD, medication use, and hypertension awareness.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Methodology</b>	<b>1</b>
2.1	Weighting the Survey Data . . . . .	1
2.1.1	Multi-year Adjustment . . . . .	1
2.1.2	Adjust weight for blood pressure and hypertension sub-population . . . . .	2
2.1.3	Multiple Imputations . . . . .	2
2.2	Logistic Regression . . . . .	2
<b>3</b>	<b>Results</b>	<b>2</b>
<b>4</b>	<b>Conclusion</b>	<b>2</b>
<b>5</b>	<b>References</b>	<b>3</b>
<b>6</b>	<b>Appendix</b>	<b>4</b>

# 1 Introduction

Effective blood pressure (BP) management is essential for reducing the risk of cardiovascular diseases. However, since 2013, there has been a concerning decline in BP control among U.S. adults with hypertension. Our study, utilizing data from the National Health and Nutrition Examination Survey (NHANES) from 1999 to 2020, investigates the potential factors contributing to this trend. The NHANES dataset, encompassing demographics, BP measurements, hypertension status, antihypertensive medication use, and co-morbidities of 59,799 U.S. adults, provides insight into the shifts in hypertension management over twenty years. We approach this analysis by handling missing data through multiple imputations and exploring relationships between various factors and BP control trends using logistic regression. The goal is to address the key elements linked to the decline in BP and enhance health outcomes for individuals with hypertension across the U.S.

## 2 Methodology

### 2.1 Weighting the Survey Data

The dataset employs ‘Full Sample 2 Year Mobile Examination Center weights.’ Each survey cycle spans two years, except for the 2017-2020 cycle, which extends from 2017 to March 2020, covering approximately 3.2 years. The weighting process involves several steps<sup>1</sup>:

- **Base Weight Calculation:** This initial step accounts for unequal selection probabilities, especially considering the over-sampling of certain demographic groups.
- **Non-Response Adjustment:** The weights are adjusted to compensate for non-response, ensuring that the sample represents the target population accurately.
- **Post-Stratification Adjustment:** Finally, the weights undergo post-stratification adjustments. This aligns the survey estimates with the U.S. civilian non-institutionalized population figures provided by the Census Bureau.

#### 2.1.1 Multi-year Adjustment

When analyzing data spanning multiple two-year NHANES cycles from 2001–2002 onwards, new multi-year weights can be calculated by dividing the existing two-year sample weights by the number of two-year cycles included in the analysis. However, due to differences in population bases, the two-year weights for the 1999-2000 cycle are not directly comparable to those for subsequent cycles. Therefore, when combining data from 1999-2000 with 2001-2002, it’s necessary to use the special 4-year sample weights provided by the NCHS, which have been adjusted for the differing reference populations.<sup>2</sup>

In our dataset, only 2-year weights are available. To correctly adjust for multi-year analysis, especially when including 1999-2000 and 2001-2002, we must source and apply the 4-year weights for these cycles. The adjustment process for combining ten survey cycles is as follows:

- For 1999-2000 and 2001-2002 cycles: Multiply the 4-year sample weights by (4/21.2).
- For the 2017-March 2020 cycle: Multiply the 2-year sample weights by (3.2/21.2), as this cycle covers approximately 3.2 years.
- For all other survey cycles: Multiply the 2-year sample weights by (2/21.2).

This method ensures that the weights are appropriately adjusted for the total span of 21.2 years covered by the ten survey cycles.

---

<sup>1</sup>Centers for Disease Control and Prevention. Nhanes tutorials - weighting module. Centers for Disease Control and Prevention. <https://www.cdc.gov/nchs/nhanes/tutorials/Weighting.aspx>

<sup>2</sup>Centers for Disease Control and Prevention. Nhanes tutorials - weighting module. Centers for Disease Control and Prevention. <https://www.cdc.gov/nchs/nhanes/tutorials/Weighting.aspx>

### 2.1.2 Adjust weight for blood pressure and hypertension sub-population

Our focus is on the sub-population with blood pressure and hypertension issues. When working with complex survey data such as NHANES, it's crucial not to exclude records from the dataset before conducting analyses. Instead, to ensure accurate variance estimates, subgroup analyses should be performed using specific functionalities within the analysis software.<sup>3</sup>

In the context of R and the survey package, we achieve this by employing the subset function. This approach allows us to correctly execute subgroup analyses on the blood pressure and hypertension population, ensuring that the variance estimates are appropriate for the complex survey design of NHANES.

### 2.1.3 Multiple Imputations

To handle missing data in our NHANES dataset, we utilize Multiple Imputation (MI) to generate multiple datasets by imputing missing values repeatedly. This method retains the advantages of single imputation, such as consistent analyses and data collector's knowledge, while also accurately reflecting uncertainty and accounting for imputation error. We implemented it using the MICE (Multivariate Imputation by Chained Equations) package in R, which allows for flexible and efficient imputation of missing values. This ensures that the imputed values are plausible and improves statistical efficiency.

## 2.2 Logistic Regression

In our analysis, as we are dealing with binary outcomes, specifically `bp_uncontrolled_130_80` for stage 1 hypertension (1: Yes, 0: No) and `bp_uncontrolled_140_90` for stage 2 hypertension (1: Yes, 0: No), we chose logistic regression as our statistical model.

Consider a logistic regression model with  $n$  covariates  $X_i$ ,  $i = 1, \dots, n$ , the model can be expressed as:

$$\log\left(\frac{\pi}{1-\pi}\right) = X\beta = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

- $\pi$  is the risk
- $\frac{\pi}{1-\pi}$  is the odds
- $\beta_0$  is the log odds for  $X_i' s = 0$
- $\beta_1$  is the log odds ratio per unit change of  $X_1$ , holding all other covariates fixed

## 3 Results

## 4 Conclusion

---

<sup>3</sup>Centers for Disease Control and Prevention. Nhanes tutorials - Variance Estimation module. Centers for Disease Control and Prevention. <https://wwwn.cdc.gov/nchs/nhanes/tutorials/VarianceEstimation.aspx>

## 5 References

## 6 Appendix