

ENAR 2024 DataFest Project Report

Mailman School of Public Health at Columbia University

Yi Huang, Yiying Wu

Dec 22, 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.

Keywords: blood pressure control, cardiovascular, hypertension risk, multiple imputation, NHANES data, risk analysis

Contents

1	Introduction	1
1.1	Background	1
1.2	Data Description	1
2	Methodology	3
2.1	Weighting the Survey Data	3
2.1.1	Multi-year Adjustment	3
2.1.2	Adjust weight for blood pressure and hypertension sub-population	3
2.1.3	Multiple Imputation	3
2.2	Logistic Regression	4
3	Results	4
4	Conclusion	5
5	References	7
6	Appendix	8

1 Introduction

1.1 Background

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 utilize logistic regression to explore the implications of co-morbidities. Given that aspects of the prevalence of BP control among US adults with hypertension have decreased since 2013, we hypothesize that uncontrolled indicator outcomes may differ on the measures related to personal lifestyles and comorbidities. 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.

1.2 Data Description

Table 1: Summary of Dataset by Gender

Characteristic	male, N = 28,882 [†]	female, N = 30,917 [†]
svy_year		
1999-2000	2,335 (8.1%)	2,641 (8.5%)
2001-2002	2,685 (9.3%)	2,907 (9.4%)
2003-2004	2,544 (8.8%)	2,759 (8.9%)
2005-2006	2,561 (8.9%)	2,773 (9.0%)
2007-2008	2,954 (10%)	3,041 (9.8%)
2009-2010	3,092 (11%)	3,268 (11%)
2011-2012	2,772 (9.6%)	2,843 (9.2%)
2013-2014	2,823 (9.8%)	3,101 (10%)
2015-2016	2,759 (9.6%)	2,976 (9.6%)
2017-2020	4,357 (15%)	4,608 (15%)
bp_uncontrolled_140_90	5,590 (19%)	5,492 (18%)
bp_uncontrolled_130_80	11,850 (41%)	10,178 (33%)
demo_age_years	47 (31, 64)	46 (30, 63)
demo_race		
Non-Hispanic White	12,427 (43%)	12,827 (41%)
Hispanic/Asian/Other	10,084 (35%)	11,178 (36%)
Non-Hispanic Black	6,371 (22%)	6,912 (22%)
cc_bmi		
<25	8,622 (31%)	9,742 (32%)
25 to <30	10,541 (37%)	8,545 (28%)
30+	8,948 (32%)	11,823 (39%)
Unknown	771	807
cc_diabetes	4,018 (14%)	3,764 (12%)
cc_ckd	4,663 (16%)	5,316 (17%)
cc_cvd_any	3,347 (12%)	2,476 (8.0%)
bp_med_use	6,751 (24%)	7,939 (26%)
Unknown	198	94
htn_aware	9,336 (33%)	10,311 (33%)
Unknown	183	85

¹n (%); Median (IQR)

Combined Table: Summary of Dataset by Blood Pressure Control Status

Characteristic	BP 140/90		BP 130/80	
	Controlled, N = 48,717 ¹	Uncontrolled, N = 11,082 ¹	Controlled, N = 37,771 ¹	Uncontrolled, N = 11,082 ¹
svy__year				
1999-2000	3,831 (7.9%)	1,145 (10%)	2,864 (7.6%)	1,145 (10%)
2001-2002	4,485 (9.2%)	1,107 (10.0%)	3,409 (9.0%)	1,107 (10.0%)
2003-2004	4,260 (8.7%)	1,043 (9.4%)	3,335 (8.8%)	1,043 (9.4%)
2005-2006	4,450 (9.1%)	884 (8.0%)	3,547 (9.4%)	884 (8.0%)
2007-2008	4,863 (10.0%)	1,132 (10%)	3,800 (10%)	1,132 (10%)
2009-2010	5,310 (11%)	1,050 (9.5%)	4,151 (11%)	1,050 (9.5%)
2011-2012	4,679 (9.6%)	936 (8.4%)	3,576 (9.5%)	936 (8.4%)
2013-2014	4,985 (10%)	939 (8.5%)	3,924 (10%)	939 (8.5%)
2015-2016	4,671 (9.6%)	1,064 (9.6%)	3,551 (9.4%)	1,064 (9.6%)
2017-2020	7,183 (15%)	1,782 (16%)	5,614 (15%)	1,782 (16%)
demo__age__years	42 (28, 59)	64 (52, 75)	39 (26, 56)	64 (52, 75)
demo__race				
Non-Hispanic White	20,562 (42%)	4,692 (42%)	15,714 (42%)	4,692 (42%)
Hispanic/Asian/Other	17,984 (37%)	3,278 (30%)	14,475 (38%)	3,278 (30%)
Non-Hispanic Black	10,171 (21%)	3,112 (28%)	7,582 (20%)	3,112 (28%)
gender				
male	23,292 (48%)	5,590 (50%)	17,032 (45%)	5,590 (50%)
female	25,425 (52%)	5,492 (50%)	20,739 (55%)	5,492 (50%)
cc__bmi				
<25	15,719 (33%)	2,645 (25%)	13,199 (36%)	2,645 (25%)
25 to <30	15,435 (33%)	3,651 (34%)	11,854 (32%)	3,651 (34%)
30+	16,324 (34%)	4,447 (41%)	11,708 (32%)	4,447 (41%)
Unknown	1,239	339	1,010	339
cc__diabetes	5,332 (11%)	2,450 (22%)	3,602 (9.5%)	2,450 (22%)
cc__ckd	6,253 (13%)	3,726 (34%)	4,352 (12%)	3,726 (34%)
cc__cvd__any	3,874 (8.0%)	1,949 (18%)	2,752 (7.3%)	1,949 (18%)
bp__med__use	9,327 (19%)	5,363 (49%)	6,028 (16%)	5,363 (49%)
Unknown	254	38	212	38
htn__aware	12,677 (26%)	6,970 (63%)	8,191 (22%)	6,970 (63%)
Unknown	233	35	199	35

¹n (%); Median (IQR)

The data show an increase in controlled blood pressure, peaking in 2017-2020. Older age correlates with uncontrolled blood pressure, and racial disparities are evident, especially among Non-Hispanic Blacks. Women generally have better control than men. Higher BMI, diabetes, and chronic kidney disease are more common in the uncontrolled group, indicating management challenges. Despite medication use, many have uncontrolled blood pressure, suggesting issues with treatment efficacy or compliance. Higher hypertension awareness doesn't always lead to control, emphasizing the need for effective management strategies. These insights underscore the complexity of blood pressure management across various demographic and health factors.

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 steps¹:

- **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.²

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.

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.³

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 Imputation

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,

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

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

³Centers 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>

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

Given two binary outcomes, `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. Suppose there are n covariates $X_i, i = 1, \dots, n$, the model can be expressed as:

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

- π is the risk
- $\frac{\pi}{1-\pi}$ is the odds
- β_0 is the log odds for $X'_i s = 0$
- β_i is the log odds ratio per unit change of X_i , holding all other covariates fixed

In our logistic regression analysis for survey data, we have integrated multiple imputation and design effects. This approach started with performing multiple imputations using MICE to create complete datasets for missing data. Each imputed dataset was then individually analyzed with logistic regression models, incorporating essential survey design elements like stratification and weights via the survey package. Lastly, aggregating these results to derive final estimates.

3 Results

In Table 2 and Figure 1, age exhibited significance in both stages of hypertension outcomes, with each additional year of age associated with an increased odds ratio for worsening blood pressure control. Non-Hispanic Black individuals displayed significantly higher odds ratios for both stages (1.59 with a 95% CI of 1.5-1.69 for stage 1, and 1.96 with a 95% CI of 1.82- 2.11 for stage 2), suggesting racial disparities in hypertension management. Additionally, individuals with a BMI of 30 or higher had elevated odds ratios for worsening blood pressure control compared to those with a BMI less than 25, highlighting the importance of weight management.

Table 2: Summary of Logistic Regression Models

Variable	SBP >130 and DBP >80		SBP >140 and DBP >90	
	OR (95% CI)	P-value	OR (95% CI)	P-value
Survey year				
2001-2002	0.89 (0.80, 1.00)	0.0500	0.87 (0.76, 1.00)	0.0548
2003-2004	0.76 (0.68, 0.86)	<0.0001	0.83 (0.72, 0.95)	0.0084
2004-2006	0.70 (0.62, 0.78)	<0.0001	0.73 (0.63, 0.84)	<0.0001
2007-2008	0.62 (0.55, 0.70)	<0.0001	0.65 (0.56, 0.74)	<0.0001
2009-2010	0.56 (0.50, 0.63)	<0.0001	0.54 (0.47, 0.62)	<0.0001
2011-2012	0.60 (0.53, 0.68)	<0.0001	0.55 (0.48, 0.65)	<0.0001
2013-2014	0.50 (0.45, 0.57)	<0.0001	0.51 (0.44, 0.59)	<0.0001
2015-2016	0.59 (0.52, 0.66)	<0.0001	0.60 (0.52, 0.69)	<0.0001
2017-2020	0.58 (0.52, 0.65)	<0.0001	0.62 (0.54, 0.72)	<0.0001
Age (Years)	1.04 (1.04, 1.04)	<0.0001	1.05 (1.05, 1.06)	<0.0001
Race				
Race Hispanic/Asian/Other	1.08 (1.02, 1.15)	0.0095	1.25 (1.15, 1.35)	<0.0001
Race Non-Hispanic Black	1.59 (1.50, 1.69)	<0.0001	1.96 (1.82, 2.11)	<0.0001
Gender				
Female	0.64 (0.61, 0.67)	<0.0001	0.82 (0.77, 0.88)	<0.0001
BMI				
25 to <30	1.23 (1.15, 1.31)	<0.0001	1.06 (0.97, 1.15)	0.2209
30+	1.72 (1.61, 1.84)	<0.0001	1.29 (1.18, 1.40)	<0.0001
Diabetes	0.91 (0.83, 0.99)	0.0295	0.94 (0.86, 1.03)	0.2077
CKD	1.40 (1.30, 1.51)	<0.0001	1.59 (1.46, 1.72)	<0.0001
CVD	0.69 (0.63, 0.76)	<0.0001	0.74 (0.67, 0.81)	<0.0001
Medication Use	0.57 (0.51, 0.63)	<0.0001	0.54 (0.48, 0.60)	<0.0001
Hypertension Awareness	3.37 (3.08, 3.70)	<0.0001	4.07 (3.67, 4.51)	<0.0001
SBP: Systolic Blood Pressure; DBP: Diastolic Blood Pressure Stage 1 Hypertension: (SBP \geq 130 mm Hg or DBP \geq 80 mm Hg) Stage 2 Hypertension: (SBP \geq 140 mm Hg or DBP \geq 90 mm Hg)				

Source: Centers for Disease Control and Prevention (CDC). [High Blood Pressure Facts](#).

4 Conclusion

From our analysis, we observe survey year, age, race, gender, BMI, diabetes, chronic kidney disease, cardiovascular disease, medication use, and hypertension awareness are statistically significant in our model. This supports our first hypothesis that lifestyle and comorbidities might impact the risk of hypertension. Individuals with older age, non-Hispanic Black ethnicity, obesity, and chronic kidney disease were associated with higher odds of worsening blood pressure control. Furthermore, we observed a significant trend over

the years, indicating that blood pressure control has improved in more recent survey years compared to the baseline year 1999-2000. This trend continued even after adjusting for other covariates in the model. Individuals with hypertension awareness had significantly higher odds of worsening blood pressure control because individuals with hypertension awareness typically have a higher likelihood of being diagnosed and treated for hypertension.

There are some limitations in the current analysis, such as the absence of a variable selection process and trend analysis. Our analysis relied on logistic regression with a specific set of predictor variables, which did not take into account all the relevant predictors from the dataset. By using Lasso for variable selection, we could avoid potential overfitting issues in our analysis. Our main objective is to investigate the potential causes of worsening BP control among US adults with hypertension over the past decade. A time series model would be more effective than logistic regression in accounting for temporal dependencies and capturing trends and long-term changes.

5 References

“Facts about Hypertension.” Centers for Disease Control and Prevention, 6 July 2023. URL <http://www.cdc.gov/bloodpressure/facts.htm/> [Accessed 22 Dec. 2023.]

“Nhanes tutorials - Variance Estimation module”. Centers for Disease Control and Prevention, URL <https://wwwn.cdc.gov/nchs/nhanes/tutorials/VarianceEstimation.aspx> [Accessed 22 Dec. 2023.]

“Nhanes tutorials - weighting module”. Centers for Disease Control and Prevention, URL <https://wwwn.cdc.gov/nchs/nhanes/tutorials/Weighting.aspx> [Accessed 22 Dec. 2023.]

6 Appendix