What Caused the Decline in Blood Pressure Control After 2013?

An Analysis Report Based on NHANES Survey Data

Xianjie Que

Biomedical Data Mining – Final Project

Washington University in St. Louis

May 9, 2025

1. Introduction and Objective

Effective blood pressure (BP) control is essential for reducing cardiovascular risk, yet national data show a decline among U.S. adults with hypertension since 2013. This study investigates potential factors behind this trend using NHANES data (n = 26,757; 160 variables). The outcome is a binary indicator of BP control ("*bp_control_accaha*").

2. Data Preprocessing

2.1 Phase Classification

To capture temporal trends, a categorical variable named "*phase*" was created to distinguish between two periods: the rising phase (pre-2013) and the falling phase (post-2013). The rising phase served as the reference group for later analyses.

2.2 Subsetting Based on Cholesterol Data

To address high missingness in cholesterol-related variables (~58%), the dataset was stratified into two subsets: individuals with cholesterol data (n = 11,118) and individuals without cholesterol data (n = 15,639). This separation was necessary due to the structure of NHANES data: cholesterol testing was performed on a specific subsample (svy_subpop_chol == 1). Missing rates for 41 cholesterol-related variables exceeded 50%, limiting the feasibility of high-quality imputation. Given the strong clinical relationship between cholesterol and hypertension (e.g., co-occurrence with dyslipidemia, metabolic syndrome), both subsets were retained for parallel analysis. (Zidek, 2009) (Egan, 2013)

2.3 Data Cleaning

Several cleaning steps were performed:

2.3.1 Fake Missing Values: Converted placeholders like blank or "*NA*" strings into proper missing values.

2.3.2 High Collinearity: Variables with pairwise correlations ≥0.95 were identified. Redundant variables were removed in favor of those with lower missingness and stronger clinical relevance. For example, in the case of low-density lipoprotein (LDL), I retained the variable ldl_corrected and excluded LBDLDL, chol_ldl, and FriedewaldLDL.

2.3.3 Unit Standardization: Retained SI units (mmol/L) for lab variables and removed conventional units (e.g., mg/dL).

2.3.4 Avoiding Tautology: Removed actual BP measurements (e.g., BPXSY1, BPXSY2, BPXSY3, bp_sys_mean) to prevent circular reasoning in predicting BP control.

2.3.5 Near-Constant Variables: Dropped variables with negligible variance or only a single unique value.

2.3.6 Survey Design Variables: Due to extremely high collinearity among NHANES sampling design variables, these were excluded during LASSO modeling but reintroduced for logistic regression.

2.4 Imputation of Missing Values

To address missing data, I first visualized the patterns of missingness to inform the choice of imputation strategies. Variables with more than 50% missing values were excluded from the analysis to reduce bias. For the remaining variables: continuous variables were imputed using the Expectation-Maximization (EM) algorithm; categorical variables were imputed using k-Nearest Neighbors (KNN, k = 5).

2.5 Post-Imputation Collinearity Check

After imputation, another collinearity assessment was conducted, lowering the threshold to 0.80 to improve model sparsity and interpretability. Redundant variables were again filtered based on missingness and clinical relevance.

3. Statistical Analysis

3.1 LASSO with Interaction Terms

A LASSO regression model was applied to the subset with cholesterol data (n = 11,118; p = 108), including both main effects and interaction terms with the phase variable:

$$bp\_control\_accaha \sim x1 + x2 + \ldots + x1{:}phase + x2{:}phase + \ldots$$

This allowed us to test whether each variable's effect on BP control changed significantly after 2013. The rising phase was set as the reference level, so interaction terms captured shifts in associations over time.

3.2 Post-Selection Inference

3.2.1 To validate the significance of selected interaction terms, a standard survey-weighted logistic regression was conducted:

$$bp\_control\_accaha \sim x1{:}phase + x2{:}phase + x3{:}phase + \ldots$$

1) Positive coefficients indicated favorable associations with BP control.
2) Negative coefficients indicated detrimental effects.

Significant predictors and interaction terms were identified based on p-values. The significant variables selected from the subset with cholesterol data are listed in Table 1 and those from the subset without cholesterol data are in Table 2 (both ranked by p-value in descending order).

3.2.2 Trend Visualization

For selected variables, temporal trends were plotted and compared with BP control trends. Key findings included:

1) Demographics: The proportion of White individuals declined post-2013, while the proportion of Black individuals increased. (Figure 1) The proportion of adults aged 75+ also rose slightly.
2) Metabolism-Related Factors: Diabetes prevalence and markers of kidney/metabolic dysfunction increased after 2013.
3) Stable Variables: No notable changes were observed in antihypertensive medication use or drug-resistant hypertension indicators, suggesting limited contribution to the observed decline in BP control.
4) Subset-Specific Insight: In the subset without cholesterol data, the proportion of females declined post-2015—a trend not observed in the cholesterol subset. (Figure 2)

4. Conclusion

The decline in BP control among U.S. adults with hypertension since 2013 appears to be associated with:

1) Shifting Demographics: Decreased proportion of femalesand White individuals; increased proportion of Black individuals. (Sahinoz, 2021) (Everett, 2015)
2) Increased Metabolic Risk: Rising rates of diabetes, low HDL levels, and markers of kidney and metabolic dysfunction.

In contrast, variables related to cardiovascular disease, medication adherence, and treatment resistance remained relatively stable, suggesting these factors did not drive the observed decline in BP control.

5. Reflection

Working with real-world data on a self-directed research question taught me a lot.

1) I gained valuable hands-on experience throughout this project. As Professor Liu emphasized, it's crucial to understand the research question. I spent much time thinking about why BP control has declined and focused on identifying changes over time, which led me to split the data into two phases. During exploratory analysis, I learned how to detect multicollinearity, handle missing values through imputation, and determine which variables to drop or merge—skills I had never practiced before.
2) The feedback from instructors was extremely helpful. After my presentation, I received two key suggestions: use Elastic Net to address high multicollinearity and include main effects in logistic regression. These pointed out important gaps in my reasoning and knowledge. I now realize I need a better understanding of the differences among Lasso, Elastic Net, MCP, and SCAD, and when each method is most appropriate.
3) Third, the winning team's presentation served as an excellent model. I learned the importance of performing correlation analysis at the beginning to identify the most relevant variables and form a working hypothesis. Interestingly, they found that the most important predictors had no missing values, suggesting that imputation was not critical for this project—but I spending a lot of time on variable selection and imputation.
4) Lastly, I learned a lot from my classmates' projects. Some used XGBoost, while others compared Lasso, Elastic Net, and Random Forest, successfully identifying predictors similar to those found by the winning team. These approaches were both creative and inspiring.

Figures

Table 1. Significant Variables Selected from Subset with Cholesterol Data

Significant Variables Selected from Subset with Cholesterol Data

Variables with p ≤ 0.05

Variable	Estimate	Std. Error	t value	Pr(>|t|)
phaseRising:demo_age_cat45 to 64	1.0950	0.1218	8.9868	0.0000
phaseRising:demo_age_cat65 to 74	1.1911	0.1341	8.8827	0.0000
phaseRising:cc_diabetesYes	0.6312	0.1032	6.1152	0.0000
phaseRising:chol_nonhdl_5cat160 to <220 mg/dL	−0.9327	0.1539	−6.0602	0.0000
phaseRising:chol_nonhdl_5cat>= 220 mg/dL	−1.6735	0.2881	−5.8096	0.0000
phaseFalling:demo_age_cat45 to 64	0.9571	0.1697	5.6399	0.0000
phaseFalling:demo_age_cat65 to 74	1.0004	0.1897	5.2741	0.0000
phaseRising:demo_age_cat75+	0.7321	0.1520	4.8171	0.0000
phaseFalling:chol_nonhdl_5cat130 to <160 mg/dL	−0.8522	0.1784	−4.7763	0.0000
phaseRising:demo_raceNon-Hispanic White	0.6112	0.1310	4.6644	0.0000
phaseFalling:chol_nonhdl_5cat160 to <220 mg/dL	−0.8449	0.1852	−4.5627	0.0000
phaseFalling:cc_acr_gteq30Yes	−0.7068	0.1620	−4.3627	0.0000
phaseRising:htn_resistant_jnc7Yes	−1.3258	0.3444	−3.8495	0.0002
phaseRising:cc_acr_gteq30Yes	−0.3839	0.1075	−3.5698	0.0005
phaseFalling:chol_nonhdl_5cat>= 220 mg/dL	−1.2546	0.3623	−3.4629	0.0007

Table 2. Significant Variables Selected from Subset without Cholesterol Data

Significant Variables Selected from Subset with Cholesterol Data

Variables with $p \leq 0.05$

Variable	Estimate	Std. Error	t value	Pr(>|t|)
phaseRising:demo_age_cat65 to 74	1.4312	0.1496	9.5683	0.0000
phaseRising:demo_age_cat45 to 64	1.0849	0.1235	8.7870	0.0000
phaseRising:LBDSCR	0.0063	0.0008	7.7370	0.0000
phaseRising:cc_diabetesYes	0.7073	0.1032	6.8526	0.0000
phaseRising:demo_age_cat75+	0.9149	0.1358	6.7354	0.0000
phaseFalling:demo_age_cat65 to 74	1.1383	0.1718	6.6263	0.0000
phaseFalling:demo_age_cat45 to 64	0.9525	0.1591	5.9875	0.0000
phaseFalling:cc_diabetesYes	0.7502	0.1300	5.7700	0.0000
phaseRising:demo_genderWomen	0.3887	0.0905	4.2941	0.0000
phaseFalling:LBDSCR	0.0052	0.0013	4.1351	0.0001
cc_acr:phaseFalling	−0.0013	0.0003	−3.9779	0.0001
phaseFalling:demo_age_cat75+	0.6977	0.1911	3.6509	0.0004
cc_acr:phaseRising	−0.0010	0.0003	−3.6062	0.0005
phaseFalling:LBXPLTSI	0.0021	0.0006	3.2341	0.0016
phaseRising:cc_ckdYes	−0.2436	0.0876	−2.7797	0.0063

Figure 1. Trends in Race from 1999 to 2020 in Subset without Cholesterol Data

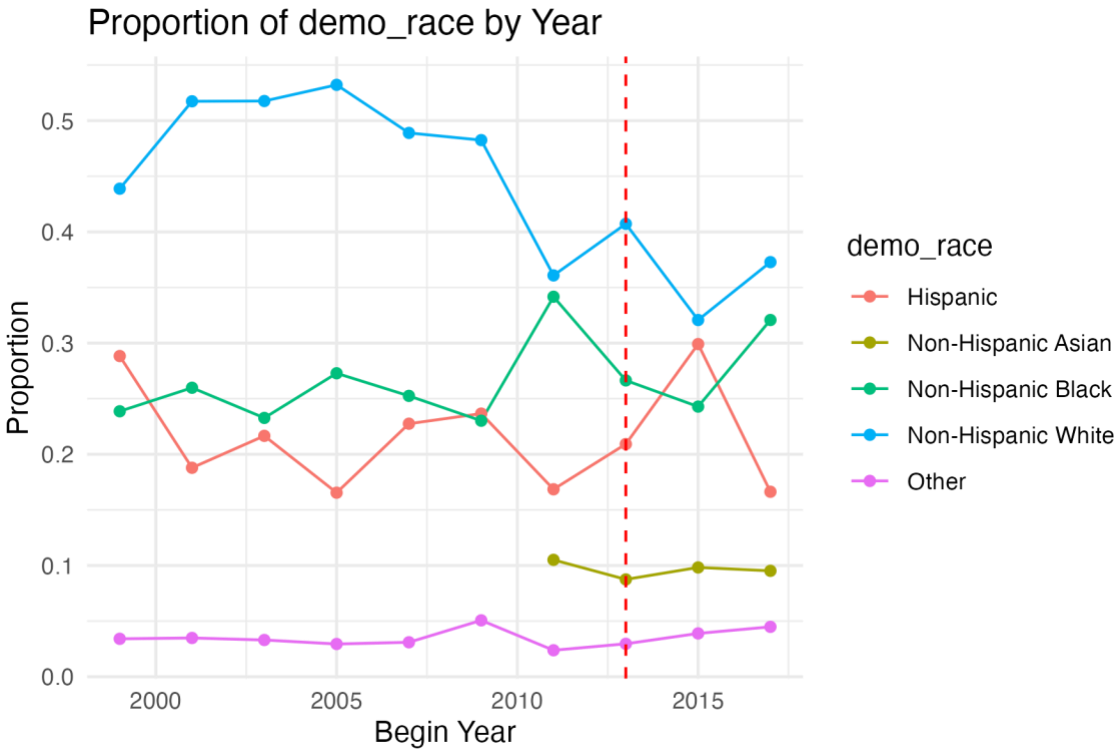
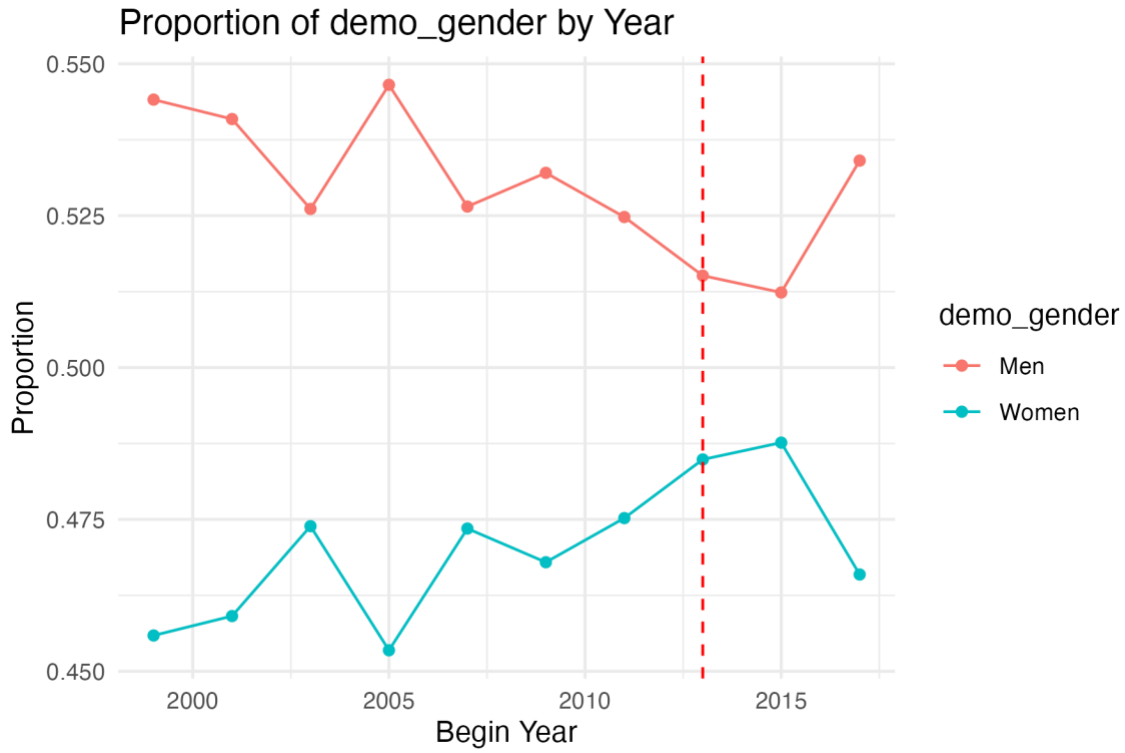


Figure 2. Trends in Gender from 1999 to 2020 in Subset without Cholesterol Data



References

1. Egan, B. M., Li, J., Qanungo, S., & Wolfman, T. E. (2013). Blood pressure and cholesterol control in hypertensive hypercholesterolemic patients: national health and nutrition examination surveys 1988-2010. *Circulation, 128(1),* 29–41. https://doi.org/10.1161/CIRCULATIONAHA.112.000500
2. Zidek, W., Naditch-Brûlé, L., Perlini, S., Farsang, C., & Kjeldsen, S. E. (2009). Blood pressure control and components of the metabolic syndrome: the GOOD survey. *Cardiovascular diabetology, 8,* 51. https://doi.org/10.1186/1475-2840-8-51
3. Everett, B., & Zajacova, A. (2015). Gender differences in hypertension and hypertension awareness among young adults. *Biodemography and social biology, 61(1),* 1–17. https://doi.org/10.1080/19485565.2014.929488
4. Sahinoz, M., Elijovich, F., Ertuglu, L. A., Ishimwe, J., Pitzer, A., Saleem, M., Mwesigwa, N., Kleyman, T. R., Laffer, C. L., & Kirabo, A. (2021). Salt Sensitivity of Blood Pressure in Blacks and Women: A Role of Inflammation, Oxidative Stress, and Epithelial Na+ Channel. *Antioxidants & redox signaling, 35(18),* 1477–1493. https://doi.org/10.1089/ars.2021.0212
5. NHANES Data. https://jhs-hwg.github.io/cardioStatsUSA/reference/nhanes_data.html (accessed April 13, 2025)

Note: All outputs, such as selected variables and plots, were saved separately in the project package.