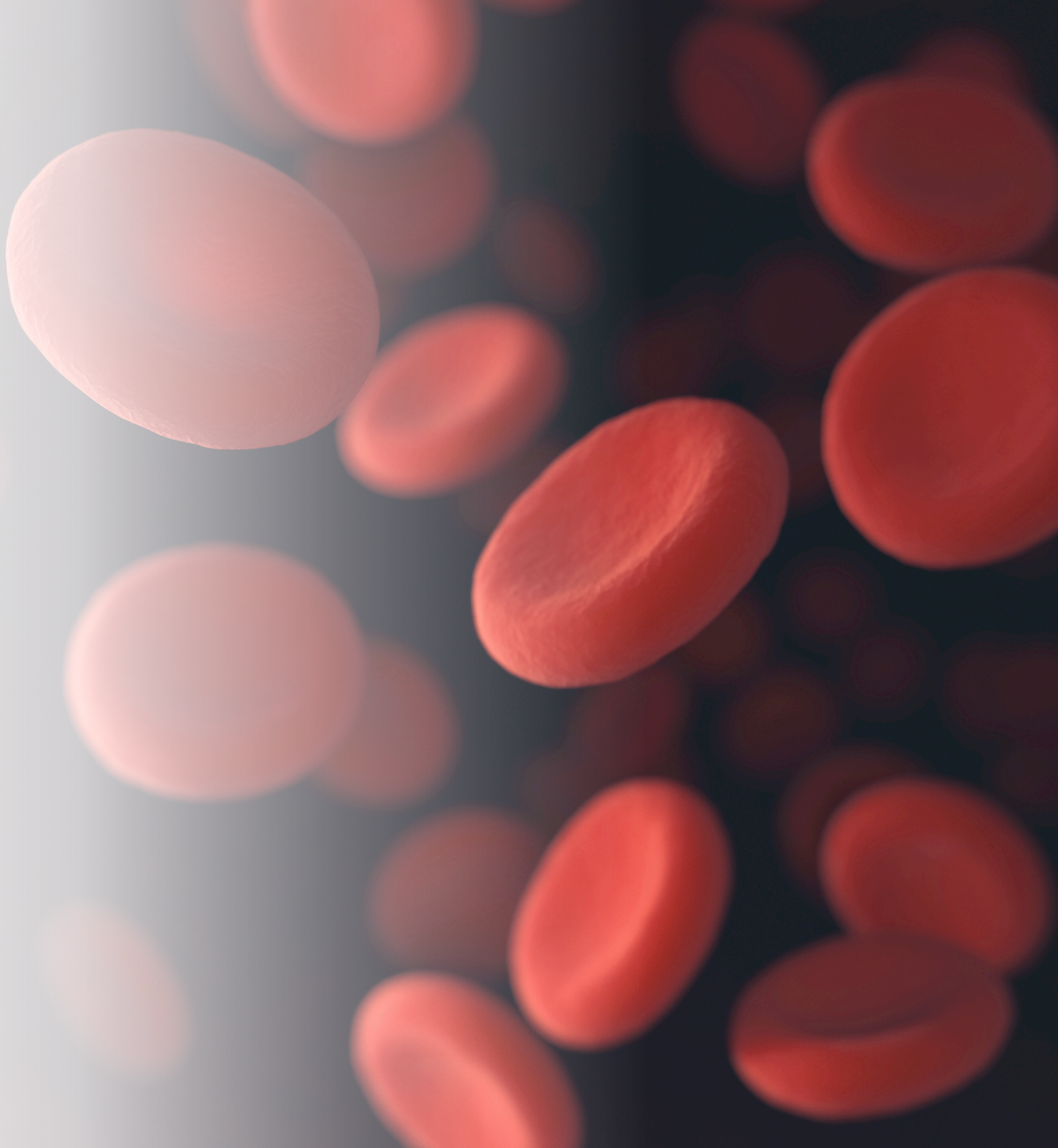
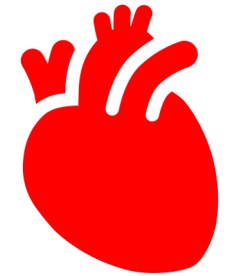
Investigating the Impact of Local Demographic and Personal Factors on Hypertension Diagnosis: A Data Science Study

Ben Pfeffer



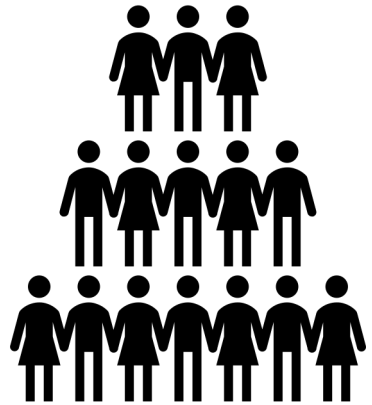
Introduction

- Uncontrolled Hypertension
  - Over 140/90
- Can lead to heart attack, stroke, aneurysm, kidney disease, etc. [1]
- Global
  - 30% of adults have hypertension, ~80% uncontrolled [2,3]
- USA
  - 50% of adults have hypertension, ~80% uncontrolled [4]
- Need increased identification of uncontrolled hypertension
- May be disparity in at-risk populations by granularity
  - Local analysis important

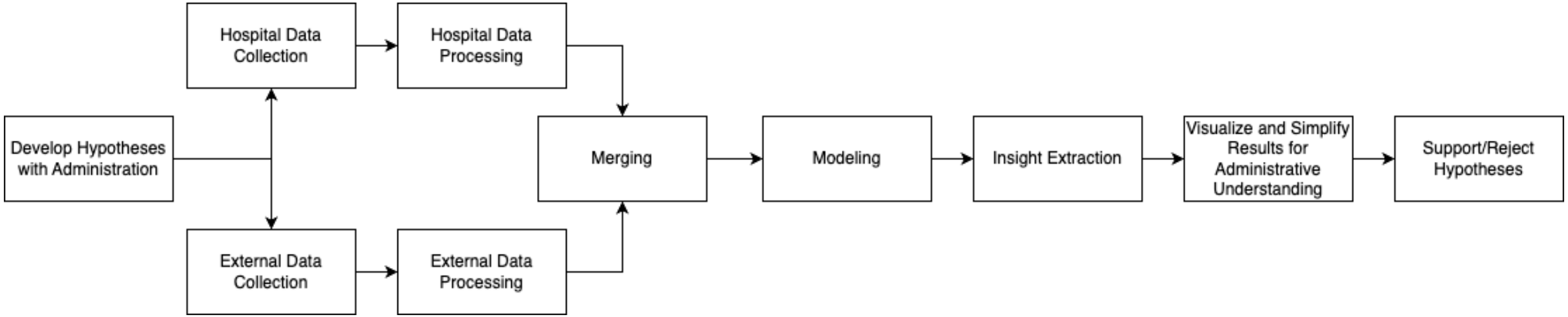


Problem

- Hospital has collected data
  - Patient information
    - Demographics
    - Habits
  - Blood pressure readings
  - HIPAA
- Need
  - Optimize care based on patient demographics
    - Aimed at *uncontrolled* hypertension
  - Ex. Provide additional screening



Process Diagram



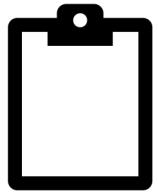
Methodology – Develop Hypotheses

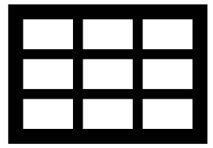
- Develop hypotheses from mass-studies
  - Black/African American = higher risk [5]
  - Lower income = higher risk [6]
  - Higher age = higher risk [7]
  - Male = higher risk [7]
  - Smoking = higher risk [8]
  - Alcohol = higher risk [8]

Methodology – Data Engineering

- Utilized 6 hospital data sets
  - 4 BP Readings March to November 2022
  - Patient Information / Demographic data
  - Questionnaire results / Personal data
- Calculated “Uncontrolled” as over 140 systolic or over 90 diastolic
- Cleaned and merged datasets

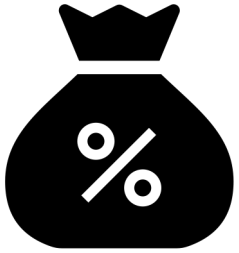
- Result:
  - People
  - Personal/demographic info
  - Diagnosis

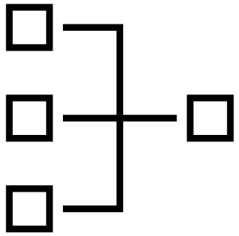




Methodology – External Data Processing

- IRS Income Tax Statistics 2020 [9]
  - Number of individuals
  - Number of returns
  - Total income
  - Grouped by zip code
  - -> Average income by zip code
- Replaced zip code in hospital data
  - Hypothesis: income affects hypertension diagnosis
  - Easier for model to understand
    - Continuous vs multiple dummy columns



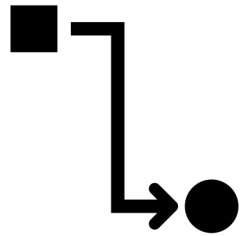


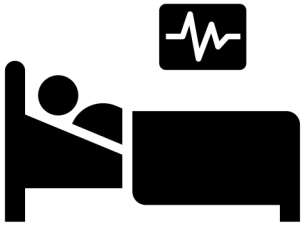
Methodology – Data Merging/Finalization

- Dummies
- 0-1 scale
- Cleaned up final data set for ML Modeling

- ~50,000 patients
- 107 columns

- ~50% of the patients had uncontrolled hypertension





Methodology – Modeling

- Classification
- Models:
  - Logistic Regression *
  - Random Forest *
  - SVC
  - KNN
- Train/Val/Test split: 0.7/0.15/0.15
- Hyperparameter optimization
- Best model -> Random Forest -> Feature importance
  - Top n features in model
  - Best n: 46

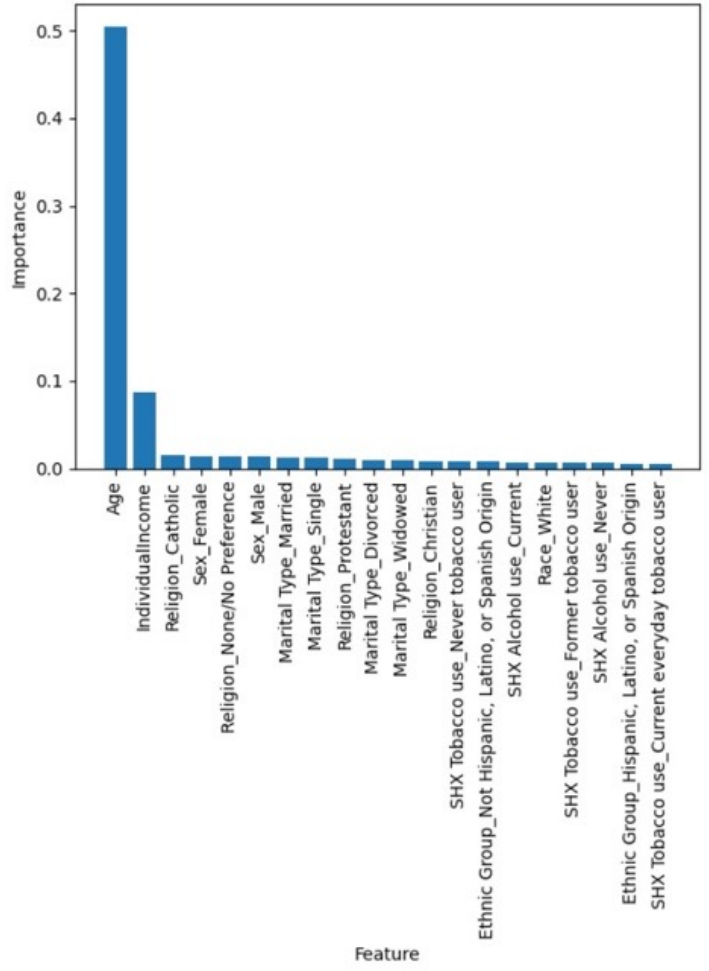
Model | Metric	Accuracy	Precision	Recall	F1 Score
Random Forest 2	?	?	?	?
Random Forest	0.7782	0.7537	0.8265	0.7884
KNN	0.7303	0.6816	0.8642	0.7621
SVC	0.7184	0.6848	0.8095	0.7420
Logistic Regression	0.7254	0.7021	0.7829	0.7403

Hyperparameter	Best Model
Max_depth	10
Max_features	log2
Min_samples_split	10
N_estimators	198

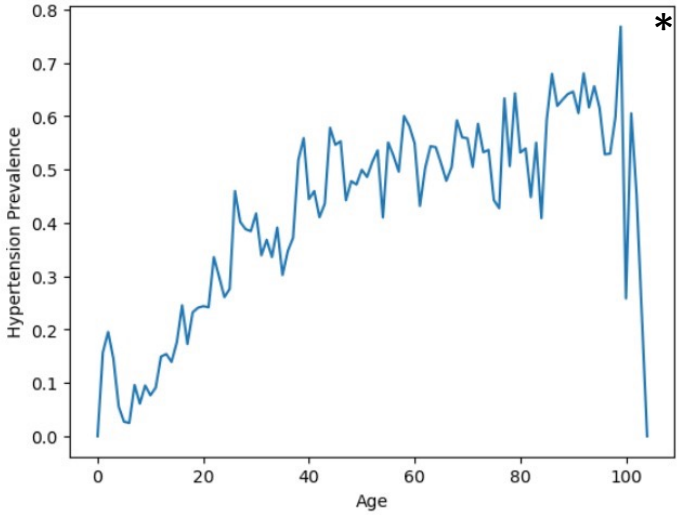
Methodology – Insight Extraction

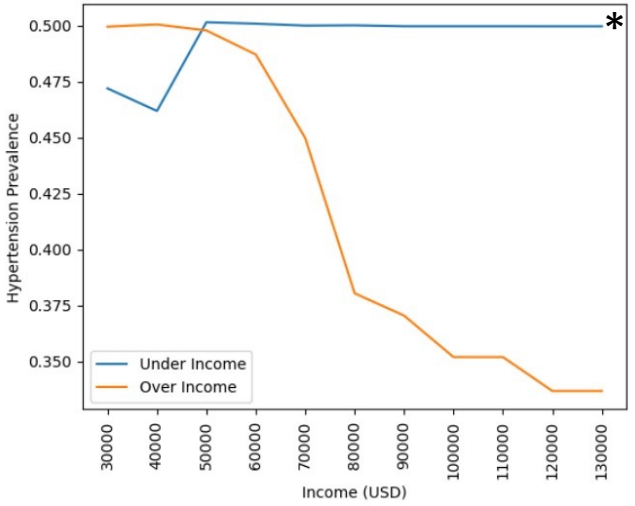
- Most impactful features extracted:
  - Age
  - Income
  - Sex
  - Smoking/Alcohol habits
- Led to accuracy of 0.8228
  - Precision: 0.7837
  - Recall: 0.8956 *
  - F1: 0.8360
- Ability to determine a health factor without any confirmed health indicators

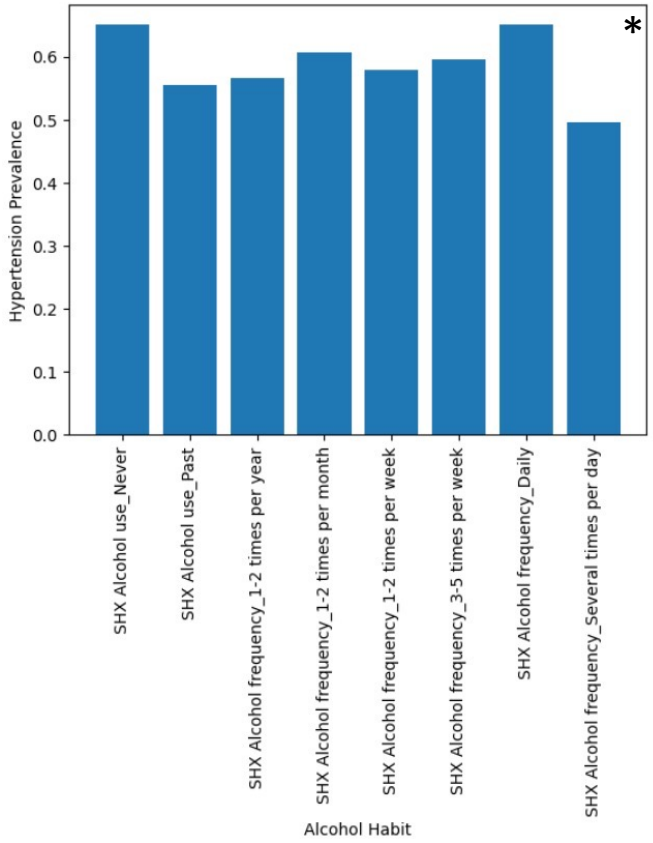
True | Pred	0	1
0	2,701	907
1	383	3,287

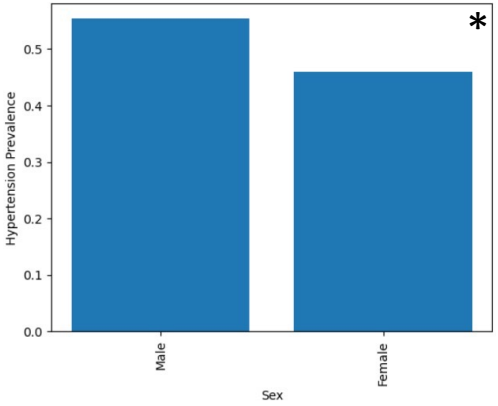


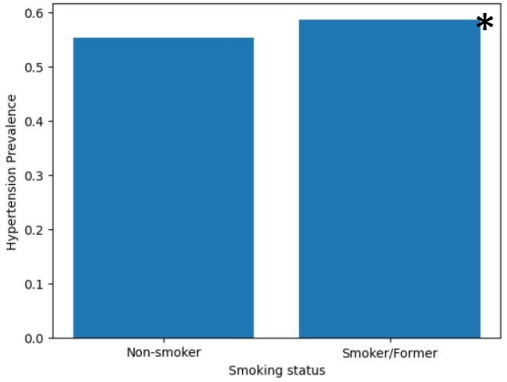
Methodology – Simplified Results

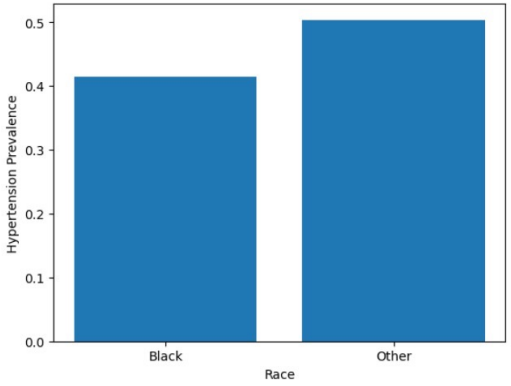












Methodology – Support/Reject Hypotheses

- Black/African American = higher risk – *Not supported* ✗
- Lower income = higher risk – *Supported* ☑
- Higher age = higher risk – *Supported* ☑
- Male = higher risk – *Supported* ☑
- Smoking = higher risk – *Supported* ☑
- Alcohol = higher risk – *Supported* ☑

Ranked by Significance

Rank	Factor
1	Age
2	Income
3	Alcohol
4	Sex
5	Smoking
6	Race

Conclusion

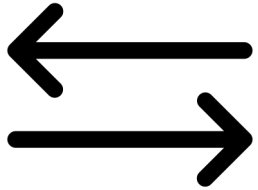
- Hypotheses supported:
  - Age is biggest factor
  - Income is an important factor
  - Being male leads to higher prevalence of UHT
  - Smoking history leads to increased risk for UHT
  - Alcohol history leads to increased risk for UHT
- Hypotheses not supported:
  - Race being Black/African American leads to highest rate of UHT
- Hospital can modify care based on results

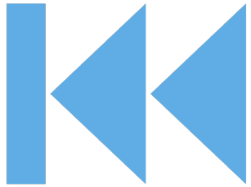




Future work

- Implement recommendation of additional screening
- Upon implementation, watch out for:
  - Unmonitored, inequivalent care can lead to severe biases
    - Need to be constantly monitored and updated
- Optimize care for other health factors
  - Diabetes
- Determine factors that impact the return to normal blood pressure
  - Will be continued in the coming weeks
- More data on the way





References

[1] “How high blood pressure can affect the body,” *Mayo Clinic*, 14-Jan-2022. [Online].Available: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868.

[2] “World Hypertension Day 2020,” *PAHO/WHO | Pan American Health Organizatio*n. [Online]. Available: https://www.paho.org/en/campaigns/world-hypertension-day-2020.

[3] “Hypertension,” *World Health Organization*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[4] Cdc, “Hypertension prevalence in the U.S.: Million hearts®,” *Centers for Disease Control and Prevention*, 22-Mar-2021. [Online]. Available: https://millionhearts.hhs.gov/data-reports/hypertension-prevalence.html.

[5] O. M. O. C.-M. Y. F. KC; “Race, ethnicity, hypertension, and heart disease: Jacc focus seminar 1/9,” *Journal of the American College of Cardiology*. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34886968/.

[6] A. Laxmaiah, I. I. Meshram, N. Arlappa, N. Balakrishna, K. M. Rao, C. G. Reddy, M. Ravindranath, S. Kumar, H. Kumar, and G. N. V. Brahmam, “Socio-Economic & Demographic Determinants of Hypertension & knowledge, practices & risk behaviour of tribals in India,” *The Indian journal of medical research*, May-2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4510771/.

[7] M. Khajedaluee, T. Hassannia, A. Rezaee, M. Ziadi, and M. Dadgarmoghaddam, “The prevalence of hypertension and its relationship with demographic factors, biochemical, and anthropometric indicators: A population-based study,” *ARYA atherosclerosis*, Nov-2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5455324/.

[8] Primatesta P;Falaschetti E;Gupta S;Marmot MG;Poulter NR; “Association between smoking and blood pressure: Evidence from the Health Survey for England,” *Hypertension (Dallas, Tex. : 1979).* [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11230269/.

[9] “Soi tax stats - individual income tax statistics - 2020 ZIP code data (SOI),” *Internal Revenue Service*. [Online]. Available: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi.