ANSWERING PUBLIC HEALTH QUESTIONS WITH NATIONAL HEALTH SURVEYS DATA ANALYSIS

Jasmine L. Chartrand, 2019

A case study of primary hypertension in NHANES 2013-2014

NHANES 2013-2014

NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

Cross-sectional survey

- Observational study from a sample of the population at a specific time

Target population:

Resident population of the United States

Sample for the interview:

10,176 persons from 30 different survey locations

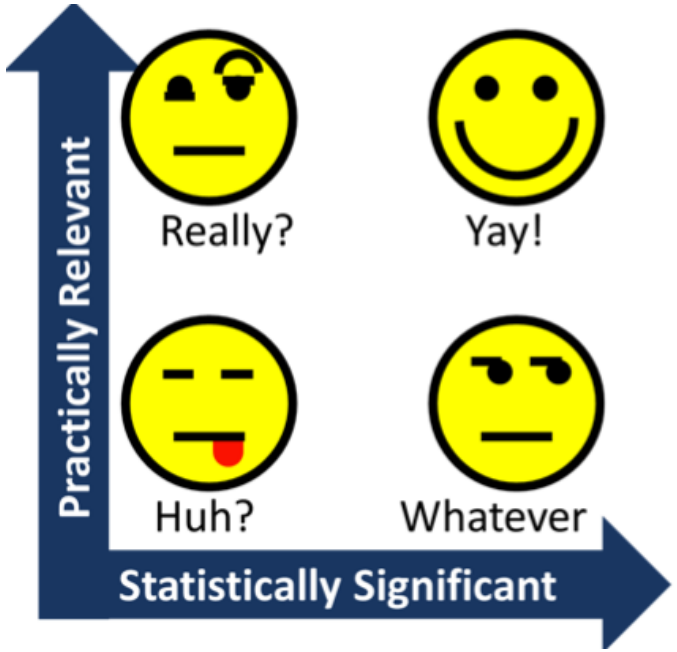
Data collected:

- Demographics, Questionnaire, Lab Results, Examination, Medication

WHY 'HYPERTENSION'?

Hypertension is the medical term for high blood pressure.

In HBP, blood applies too much force against the walls of the blood vessels.

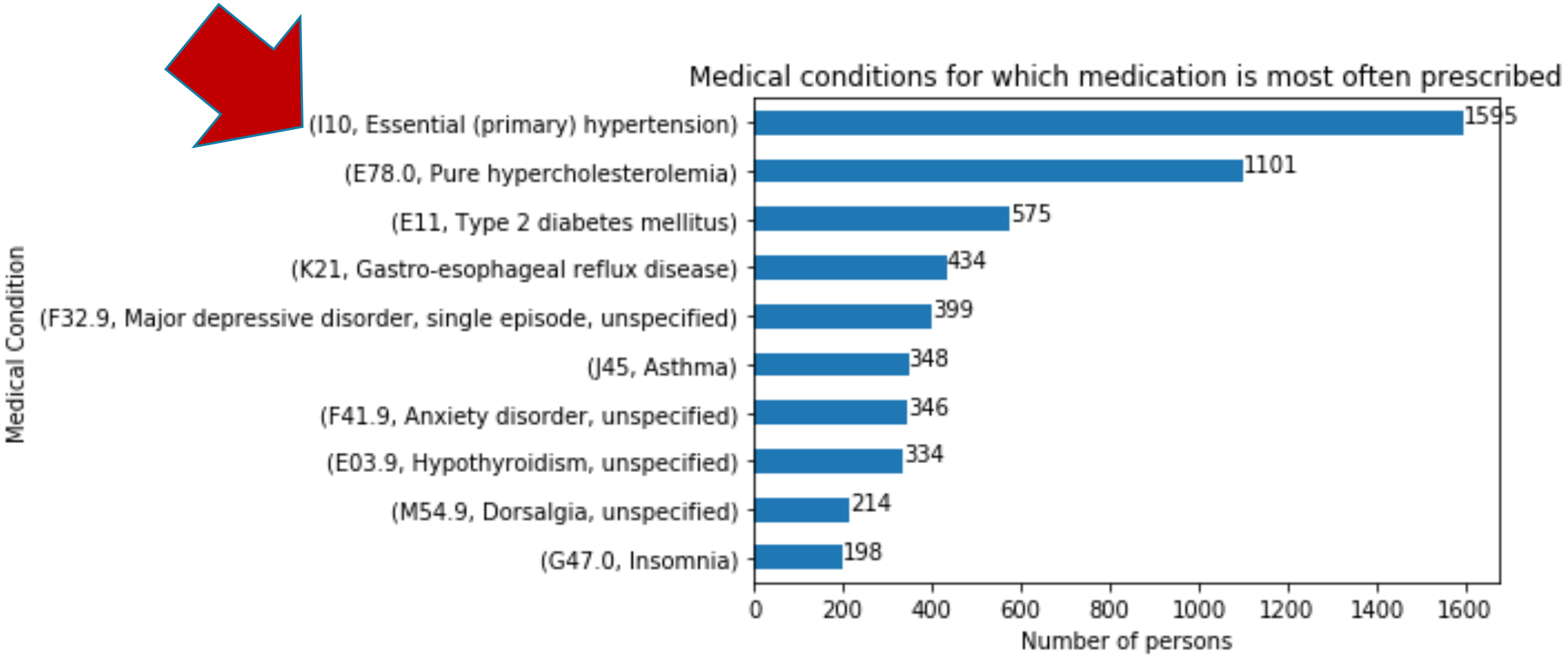


Hypertension is the most **common** primary diagnosis in the United States

Source: https://emedicine.medscape.com/article/241381-overview

WHY 'HYPERTENSION'?

HYPERTENSION IN NHANES 2013-2014



QUESTIONS

Q: What are the determinants of hypertension?

- Or what are the associated complications?
  => Simple correlation
  - (way too many missing values in each column for feature selection!)

Q: Can we identify individuals with hypertension?
=> CLASSIFICATION

Q: Can we predict the age of onset of hypertension?
=> REGRESSION

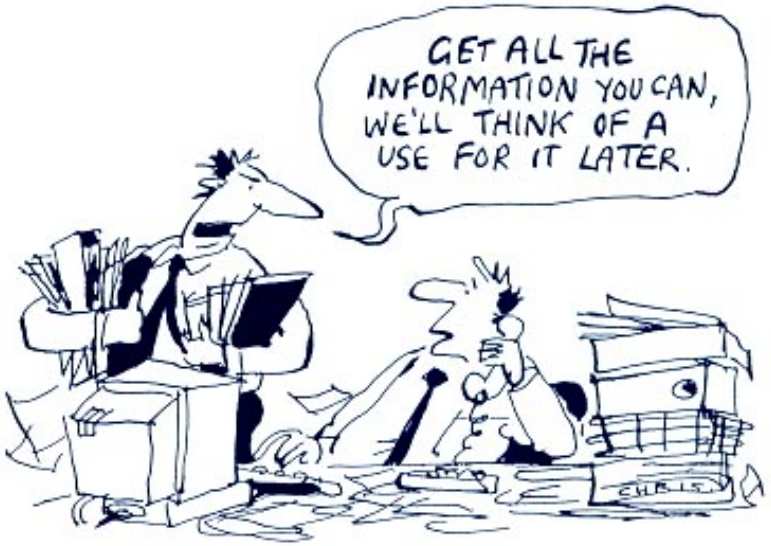
DATA ACQUISITION

Through Kaggle

- https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey
- All participants are identified by a unique sequence number 'SEQN'

The scope of this project is limited to the following files and variables

- demographics.csv
  - Age, Gender (M/F), Total Annual Household income ($), Family or Individual income ($)
- medications.csv
  - Reason for use of the medication (ICD10 medical condition code)
- questionnaire.csv
  - All questions!! => 952 variables!
    - Not all the variables are useful for our topic!



DATA CLEANING

Text values

- those variables were removed from the dataset because they were out of the scope of the project
  (i.e. 'brand of cigarettes')
- all the relevant variables were already coded as numeric values

[7, 77, 9, 99]: missing values or continuous values?

- 'Refused to answer' [value: 7, 77, 777]
- 'Don't know' [value: 9, 99, 999]
- Missing values [value: (blank) or .]

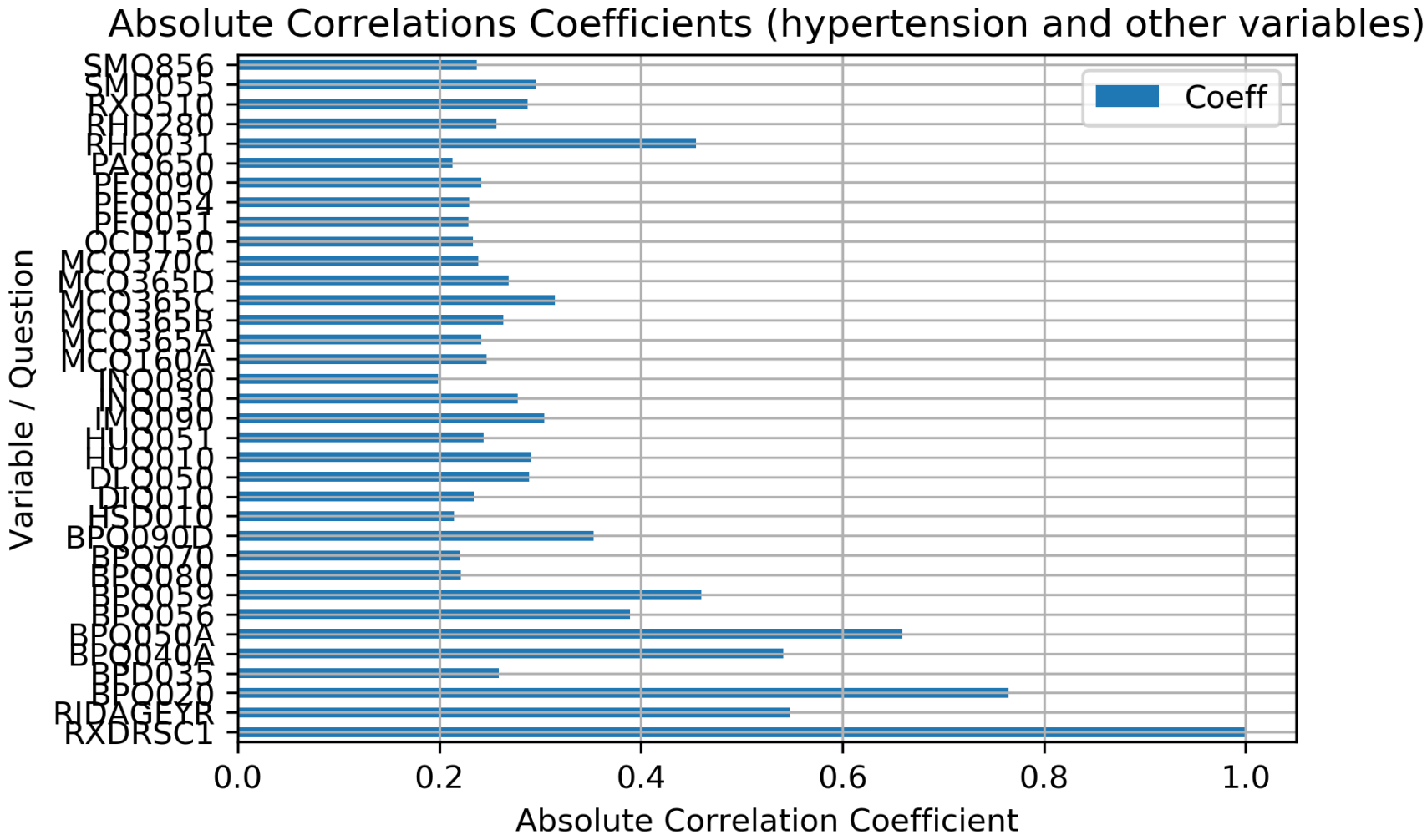
Statistically significant enough variables

- Columns where count > 310 (confidence=0.95, n=1595)

Variables with std() = 0 (variance is 0)

- i.e. 'Do you speak English at home?' (yes: 1, missing value otherwise)

Q: WHAT ARE THE QUESTIONNAIRE VARIABLES CORRELATED WITH HYPERTENSION?



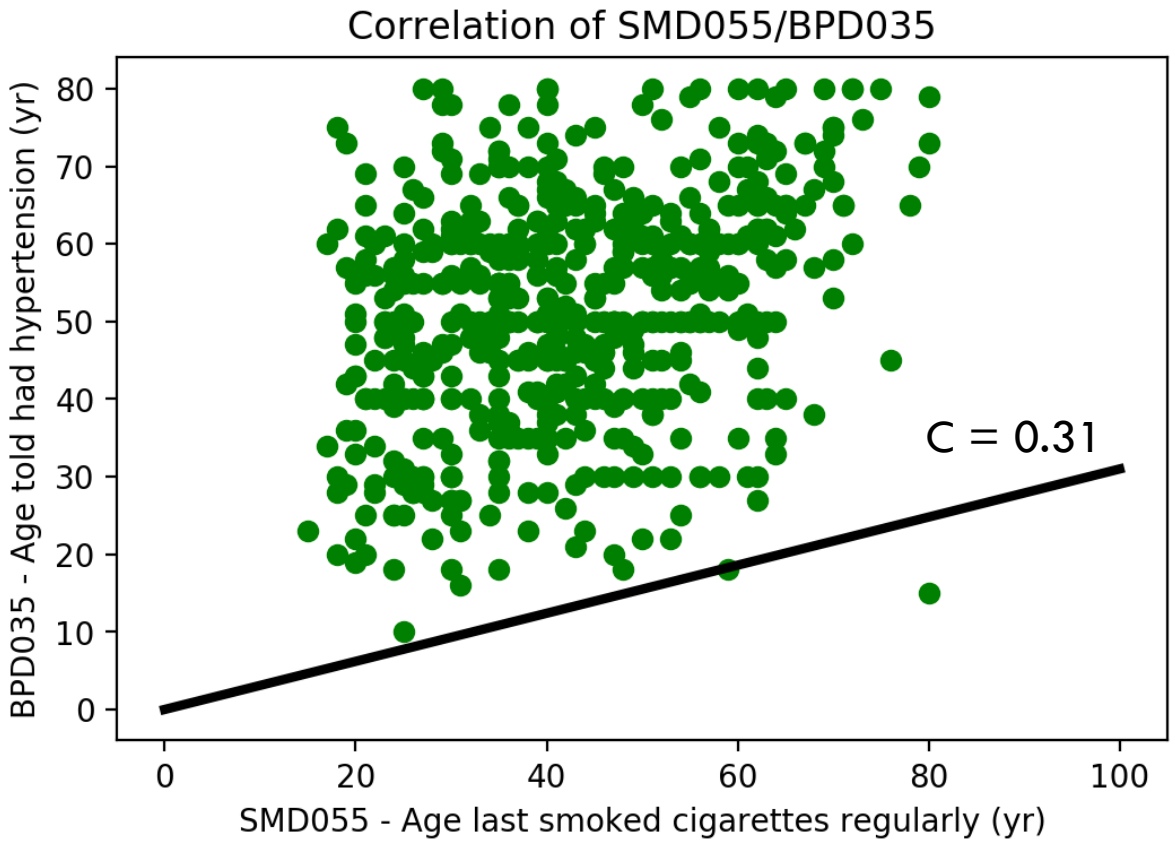
Variables with absolute correlation coeff > 0.2 with hypertension

Q: WHAT ARE THE QUESTIONNAIRE VARIABLES CORRELATED WITH HYPERTENSION?

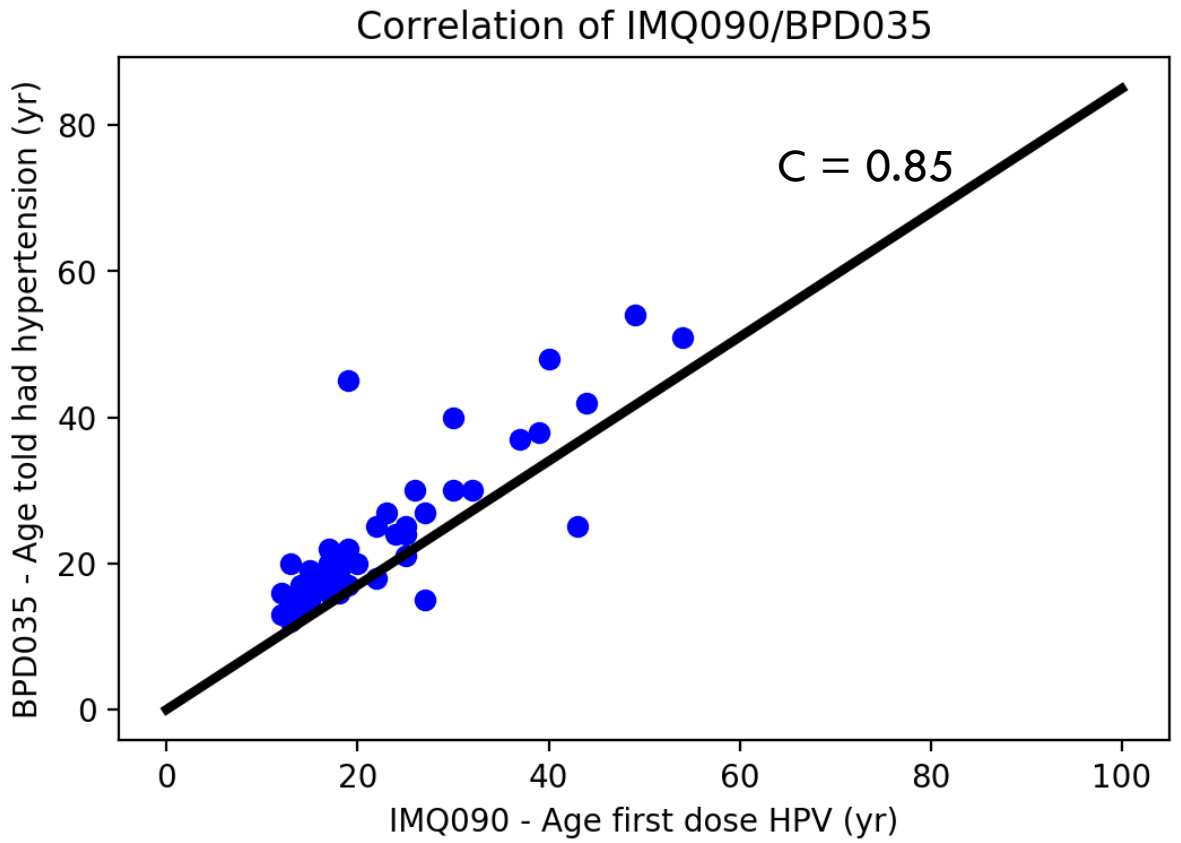
	# non-null values	
RXDRSC1	10175	Reason for use of medication
BPQ020	6464	Ever told you had high blood pressure
BPD035	2127	Age told had hypertension
BPQ040A	2174	Taking prescription for hypertension
BPQ050A	1815	Now taking prescribed medicine for HBP
BPQ056	6464	Take blood pressure at home last 12 mos?
BPQ059	6464	Doctor tell you to take BP at home?
BPQ080	6464	Doctor told you - high cholesterol level
BPQ070	4620	When blood cholesterol last checked
BPQ090D	4620	Told to take prescriptn for cholesterol
HSD010	6467	General health condition
DIQ010	9769	Have serious difficulty hearing?
DLQ050	8780	Have serious difficulty walking ?
HUQ010	10175	General health condition
HUQ051	10164	# times receive healthcare over past year
IMQ090	796	Income from Supplemental Security Income
INQ030	10052	Income from Social Security or RR
INQ080	10052	Income from retirement/survivor pension
MCQ160A	5769	Doctor ever said you had arthritis
MCQ365A	6464	Doctor told you to lose weight
MCQ365B	6464	Doctor told you to exercise
MCQ365C	6464	Doctor told you to reduce salt in diet
MCQ365D	6464	Doctor told you to reduce fat/calories
MCQ370C	6464	Are you now reducing salt in diet
OCD150	6459	Type of work done last week
PFQ051	5769	Limited in amount of work you can do
PFQ054	5769	Need special equipment to walk
PFQ090	5769	Require special healthcare equipment
PAQ650	7147	Vigorous recreational activities
RHQ031	3256	Had regular periods in past 12 months
RHD280	2620	Had a hysterectomy?
RXQ510	3815	Dr told to take daily low-dose aspirin?
SMD055	1203	Age last smoked cigarettes regularly
SMQ856	6113	Last 7-d worked at job not at home?

EXPLORING CORRELATIONS

WITH AGE-RELATED VARIABLES



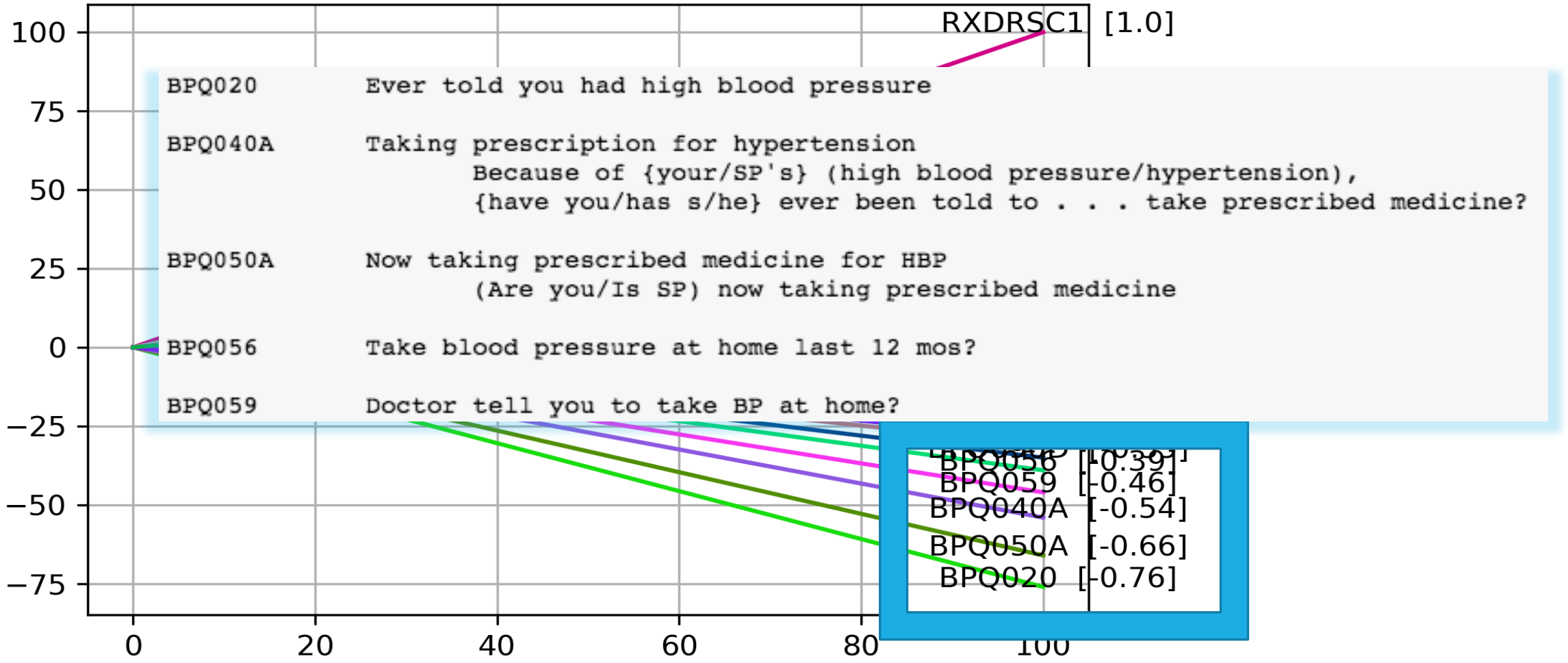
N = 576



N = 46

EXPLORING CORRELATIONS

WITH BLOOD PRESSURE VARIABLES



SOME HEALTH DATA ISSUES

Compliance

- Does the person follows the doctor's recommendations?

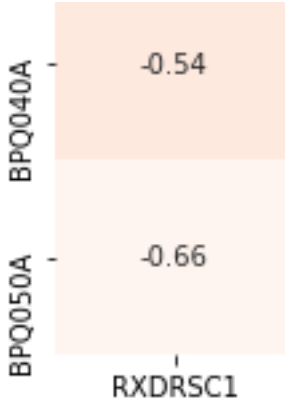
Self-reported items

- Does the person respond truthfully?

Observations at a very specific time

- Chronic VS Episodic disease

Correlation Coefficients




```
RXDRSC1  Reason for use of medication
            Use of medication for hypertension

BPQ040A  Taking prescription for hypertension
            Because of {your/SP's} (high blood pressure/hypertension),
            {have you/has s/he} ever been told to . . .
            take prescribed medicine?

BPQ050A  Now taking prescribed medicine for HBP
            (Are you/Is SP) now taking prescribed medicine
```

Q: CAN WE IDENTIFY INDIVIDUALS WITH HYPERTENSION?

Assuming we do classify an individual as having/not having hypertension only based on this target

	# non-null values	description	corr coeff
RXDRSC1	10175	Reason for use of medication	1.000000

We built our model on these arbitrary variables

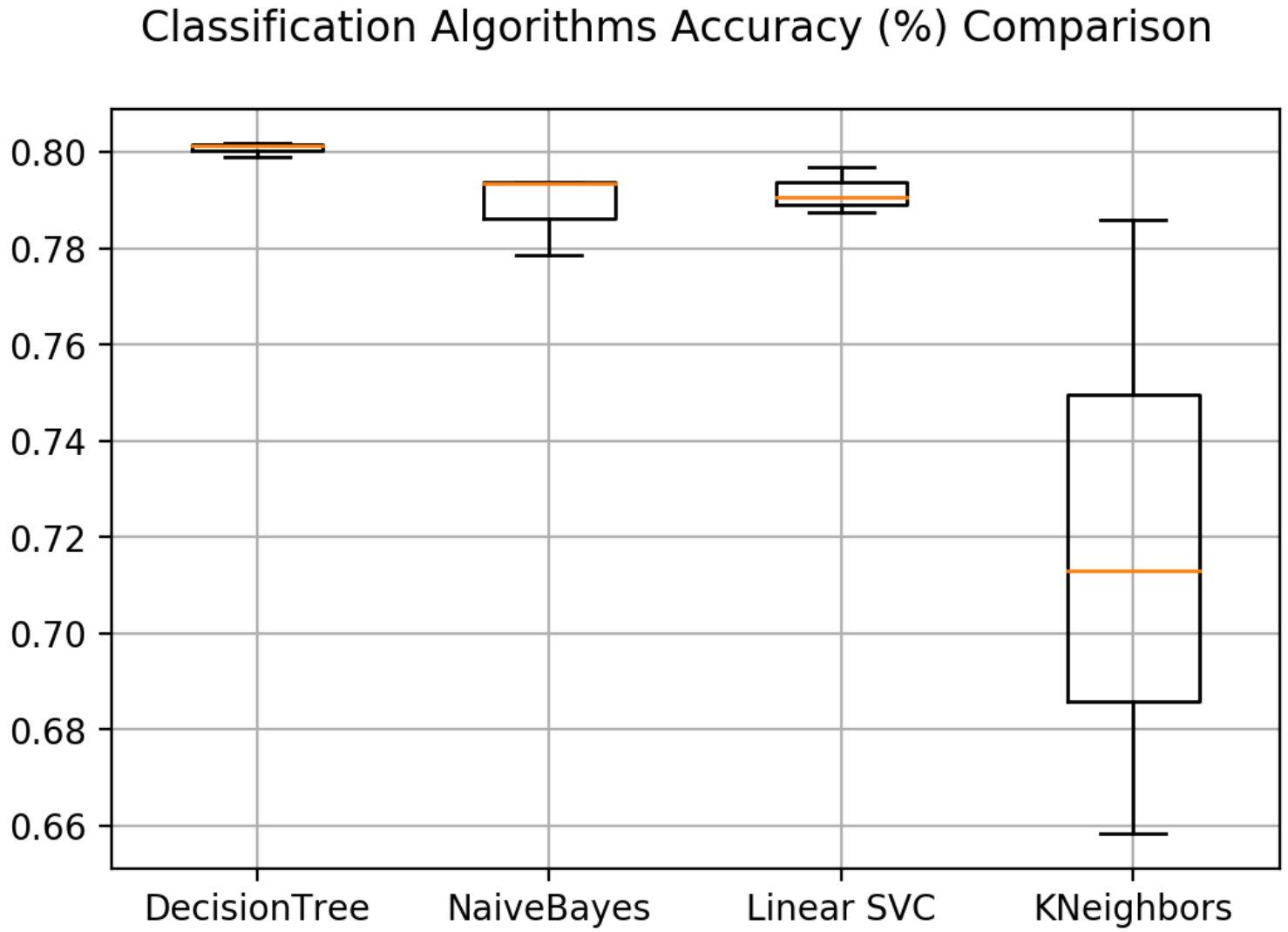
- These questionnaire variables are not too obviously related to blood pressure, and they are broad enough

	# non-null values	description	corr coeff
BPQ080	6464	Doctor told you - high cholesterol level	-0.221391
HSD010	6467	General health condition	0.214602
INQ030	10052	Income from Social Security or RR	-0.277743
INQ080	10052	Income from retirement/survivor pension	-0.198916
PAQ650	7147	Vigorous recreational activities	0.212924

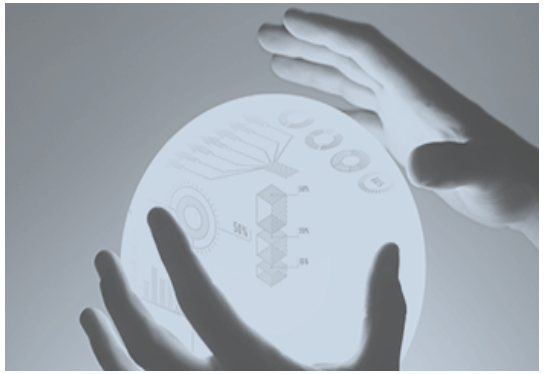
Q: CAN WE IDENTIFY INDIVIDUALS WITH HYPERTENSION?

(CONT'D)

```
DecisionTree accuracy:
    0.8007
NaiveBayes accuracy:
    0.7887
Linear SVC accuracy:
    0.7917
KNeighbors accuracy:
    0.7192
```



Q: CAN WE PREDICT THE AGE OF HYPERTENSION DIAGNOSIS?



- Target variable for regression

```
         # non-null values
BPD035     2127  Age told had hypertension
```

We built our regression model on these arbitrary variables

- These questionnaire variables may give some indications of age, and they are correlated to hypertension

```
         # non-null values
RXDRSC1    10175 Reason for use of medication
RIDAGEYR   10175 Age in years, at the time of the screening interview
BPQ020     6464  Ever told you had high blood pressure
SMD055     1203  Age last smoked cigarettes regularly
HUQ010     10175 General health condition
```

Q: CAN WE PREDICT THE AGE OF HYPERTENSION DIAGNOSIS?

(CONT'D)

Difficult to measure the performance!

RMSE is 11.97, which does makes sense…

The predictions seem
to be more or less accurate and precise.

```
LinearRegression
Mean Squared Error: 143.28934352822856
Root Mean Squared Error: 11.970352690218803
```

SEQN	Actual	Predicted
80601	71.0	56.734433
78074	51.0	43.331944
81709	80.0	61.901351
78443	50.0	49.403016
74734	56.0	46.005869
77976	65.0	57.951873
78515	80.0	63.227164
74570	53.0	53.218088
77979	66.0	55.575117
78967	36.0	38.261942
81669	37.0	35.158362
82642	64.0	52.521796
83321	60.0	61.422079
74600	50.0	53.043850
77660	60.0	54.396735
78294	28.0	28.630704
80218	60.0	57.557787
76851	65.0	56.970132
81032	50.0	56.931736
79980	47.0	38.860279
80026	50.0	44.951069
...	...	...

CONCLUSION

Challenges:

Too many variables (~ 900 variables in the questionnaire dataset)

- = > did not check the description of all the variables
- Many variables were not relevant to our topic (hypertension)

Lots of missing data:

- Column had ~ 60% of their data values filled at most!
- Few overlaps among columns

Reliability of self-reported data?

- Contradictory or Incomplete information
  - (i.e. Medication is prescribed for hypertension according to the medication records, but the participant does not report taking medication for hypertension)

CONCLUSION (CONT'D)

Questions:

Q: What are the questionnaire variables correlated with hypertension?

Blood pressure-related variables, some cholesterol variables, general health, occupational difficulties (hearing/walking), physicians recommendations (weight, exercise, diet), income source, etc.

Q: Can we identify individuals with hypertension?

The accuracies of the classification models were ~ 78-80%

Q: Can we predict the age of hypertension diagnosis?

The linear regression performance in terms of root mean squared error (RMSE) was ~ 11.97 years

ANY QUESTIONS?

CONTACT

Jasmine L.-Chartrand
jasmine.leblond-chartrand@concordia.ca
https://github.com/jleblond/NHANES_data_analytics

THANK YOU!!! ☺