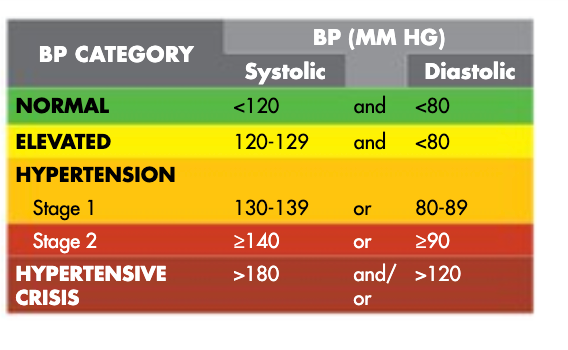
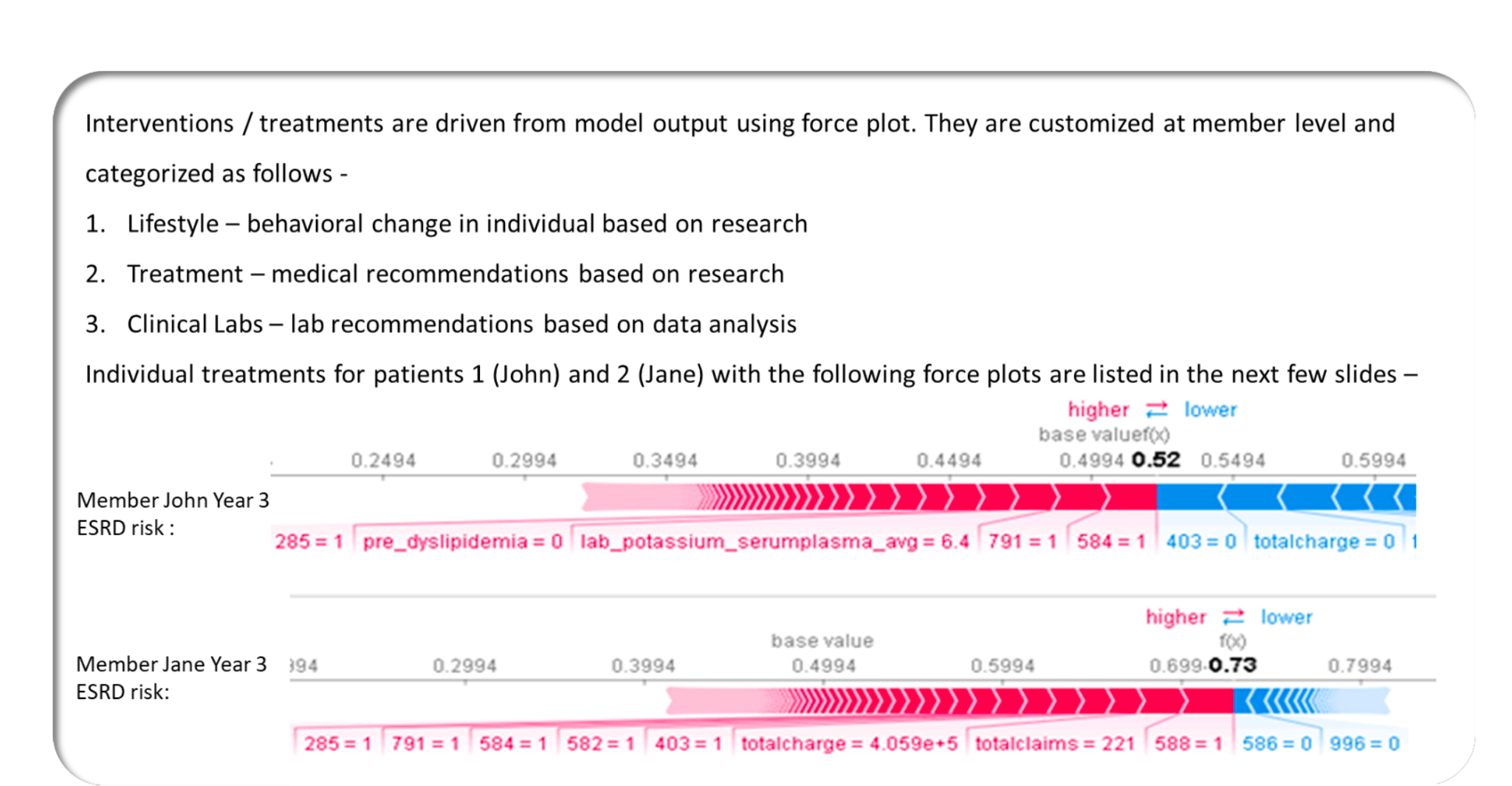
<https://wwwn.cdc.gov/nchs/nhanes/default.aspx>

Goal: identify factors that are most likely to cause hypertension (causal relationship)

* Hypertension的确认variables （3个任意一个)
  + Systolic > 130
  + Diastolic >80
  + Now taking prescribed medicine for HBP
  + Stage 1 hypertension was defined as SBP ranging 130~139 mmHg or DBP ranging 80~89 mmHg. Stage 2 hypertension was described as SBP ≥ 140 mmHg or DBP ≥ 90 mmHg[3](https://www.nature.com/articles/s41598-020-64980-8#ref-CR3).



* 想法：
  + 三种类型：

1. 没有的高血压
2. 测量中发现有高血压，之前并不知道
3. 测量中有高血压，并且知道
   * 因为已经知道自己有高血压的人可能会因此改变自己的饮食，所以建议更加关注第1，2中类型
   * 之前有一篇paper 好像提到了怎么判断哪个factor影响最大还是分析causal relation
   * Professor上课有用一个library,可以可视化看哪几个factor是降低probability(week 6 Page 31)
   * way to use SHAP:<https://github.com/slundberg/shap>
     + 

* 建立预测hypertension的模型
  + 根据上面的要求，看一共有多少人有高血压，如果数量太少，那么把加入之前年份的数据(类似于过于inbalance）
  + SMOTE 模拟真实数，相邻5个neighbor来填充
  + <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>
  + <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-106>
* 参考的维度：
  + 身体
    - Age
    - Gender
    - Race
    - BMI
    - Height
    - Weight
    - Arm, waist, hip的Circumference
  + 饮食习惯
  + 饮食(营养...)
  + 饮酒、drug、抽烟
  + 睡眠(
  + 运动(Physical Activity)
  + 其他
    - Income
    - Education
    - Citizenship
* 可以考虑的模型
* 可以尝试分开gender建立model，有paper indicate gender difference
  + Logistic Regression
  + Random Forest
  + Neural Network
    - <https://www.nature.com/articles/s41598-020-67640-z#ref-CR29>
* Measure Methods: Confusion matrix and Receiver Operating Characteristic (ROC)
* Traget AUC: 0.8 (before: AUC: 0.77 by ANN), highest is with 0.96(genetic data)
* (accuracy highly depends on datasets, so could only compare with models using the same datasets)

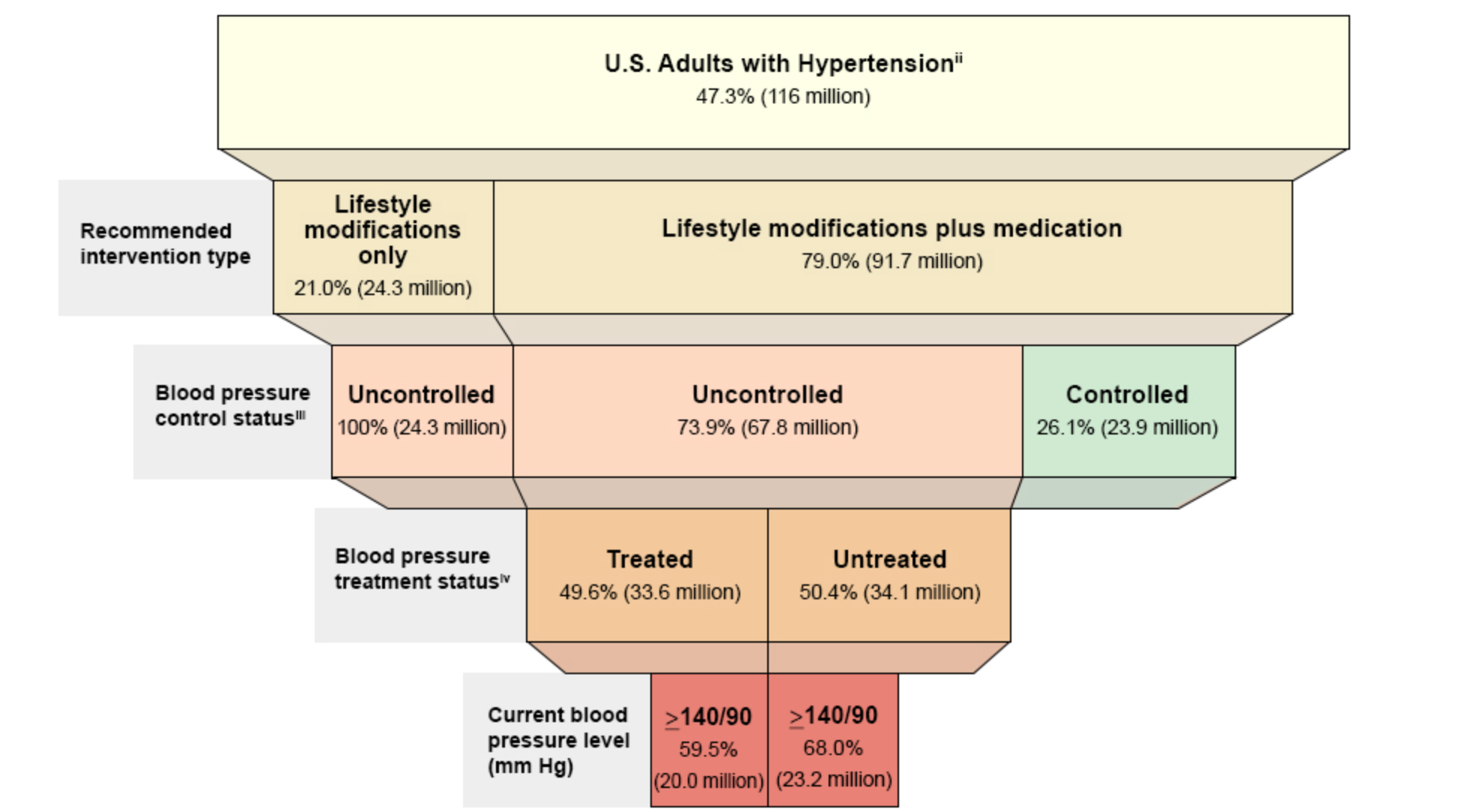
* Reference:
  + <https://github.com/AdamLiscia/Classifying-Diabetes-with-NHANES>
  + <https://thomaselove.github.io/431-2020-projectB/your2.html> (haven’t used)
* 其他
  + <https://millionhearts.hhs.gov/data-reports/hypertension-prevalence.html>

找PPT模版

模型

* Fillna –catagorical
* VIF run—Estimated 3 hours - manually review the
* Save the combine files after all data cleaning
* Rerun different models
* Select model with higher accuracy/ AUC score

SHAP—> select 5 patients



* <https://millionhearts.hhs.gov/data-reports/hypertension-prevalence-tables.html>
  + Neural Network
    - <https://www.nature.com/articles/s41598-020-67640-z#ref-CR29>
    - age, gender, BMI, systolic and diastolic blood pressure, high and low-density lipoproteins, triglycerides, cholesterol, microalbumin, and urine albumin–creatinine ratio
    - age, sex, diet, smoking and drinking habits, physical activity level and BMI index
    - age, sex, family history of hypertension, smoking habits, lipoprotein, triglyceride, uric acid, total cholesterol, and body mass index (BMI)
    - Sixteen genes, age, BMI, fasting blood sugar, hypertension medication, no history of cancer, kidney, liver or lung
    - Age, gender, race, a reason for the test, stress tests and medical history
    - age, gender, height, weight, body mass index, triglyceride, total cholesterol, low-density lipoprotein, blood urea nitrogen, uric acid, and creatinine
    - Age, gender, education level, employment, tobacco use, physical activity, adequate consumption of fruits and vegetables, abdominal obesity, history of diabetes, history of high cholesterol, and mother’s history high blood pressure

Current health status

Diabetes  
Health insurance

* Mental health
* (Occupattion)

Medical condition

* Physical activity
* Sleep disorders
* Smoking
* 有高血压，用要控制，有高血压
* 有高血压，用药空着，无高血压
* 无高血压，不知道，无高血压
* 无高血压，有高血压

Code Notes

* 2015
  + Modified
    - 'DMDHREDZ' --> 'DMDHREDU' (Household Education)
    - ALQ121 - Past 12 mo how often have alcohol drink
      * ALQ120Q - How often drink alcohol over past 12 mos per week, per month, or per year?
      * ALQ120U - # days drink alcohol per wk, mo, yr (unit)
  + Missing:
    - 'BMXHIP' (Hip) 没有
* 2013
  + Modified
    - 'CBD071' (Money spent at supermarket/grocery store)
    - → CBD070 - Money spent at supermarket/grocery store
  + Missing:
    - Sleep Disorder
      * SLQ300 - Usual sleep time on weekdays or workdays
      * SLQ310 - Usual wake time on weekdays or workdays
      * SLQ030 - How often do you snore?
      * SLQ120 - How often feel overly sleepy during day?
    - Smoke
      * SMQ890: 'smoked>1\_1\_dummy'
      * SMAQUEX2: ’Home\_Interview\_(18+ Yrs)\_1\_dummy'
  + Other:
    - SLD012 - Sleep hours (How much sleep usually get at night on weekdays or workdays?)
      * → SLD010H - How much sleep do you get (hours)?
      * SLD012 有小数
* 2007
  + race\_dummy: RIDRETH3 - Race/Hispanic origin w/ NH Asian
    - → RIDRETH1
      * 6: Non-Hispanic Asian
      * 7: Other Race - Including Multi-Racial
      * 6 + 7 都变成只有一个5: Other Race - Including Multi-Racial
  + Country of birth
    - DMDBORN4
    - → DMDBORN2
  + Diet behavior
    - Milke type:
      * 'DBD072A', 'DBQ223B', 'DBQ223C', 'DBQ223D', 'DBQ223E', 'DBQ223U'
      * → 'DBD072A', 'DBD072B', 'DBD072C', 'DBD072D', 'DBD072U'
        + 没有了soy milk, (所以之前的14变成30）
  + Smoke household
    - 没有完整数据，删除
* 2009
  + Diet behavior
    - Milke type:
      * 'DBD072A', 'DBD072B', 'DBD072C', 'DBD072D', 'DBD072U'
        + 没有了soy milk, (所以之前的14变成30）\*可考虑
      * → 'DBQ223A', 'DBQ223B', 'DBQ223C', 'DBQ223D', 'DBQ223E', 'DBQ223U'
* 2011
  + Country of birth
    - DMDBORN2
    - →DMDBORN4
* 2015
  + SLD010H

周一：

1. Based on the importance of the random forest model
2. Create heatmap based on the covariance,
3. Group top 20 important features into groups, and name them.
4. 不同的方法（shap。。。）看给出的importance feature是否一致

Find

1. If there are any possible solutions for group features.

Compare the model, if being told hypertension makes a difference.

Research if there are any papers supports our conclusion on the causal relationship

* Both healthcare and data science field

1. 删除highly correlated variables
2. based on the importance of the random forest model.

Ppt: 展示data 是否balance； Male/female； age group； race。。。

Use VIF (variance Inflation Factor): correlation with one variable to the rest variables

Iterate, drop the highest, then iterate again, drop the hightest, until the we are confidence with the remaining number of features

<https://www.statisticshowto.com/variance-inflation-factor/#:~:text=In%20general%2C%20a%20VIF%20above,like%20x%20and%20x2>.

<https://quantifyinghealth.com/vif-threshold/>

Get a high accuracy

Logistic regression

* Transform the data
* sparse, or L1-regularized,logistic regression. (Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors)

Decision Tree

Cross-validation

Models comparison

* Predicting Hospital Readmissions in the Medicare Population P2

Multiple regression

* Relative Importance Analysis: A Useful Supplement to Regression Analysis
* Mining high-dimensional administrative claims data to predict early hospital readmissions

Fillna

* https://stackoverflow.com/questions/64900801/implementing-knn-imputation-on-categorical-variables-in-an-sklearn-pipeline

Simple imputation strategies such as using the mean or median can be effective when working with univariate data. When working with multivariate data, more advanced imputation methods such as iterative imputation can lead to even better results.

https://towardsdatascience.com/iterative-imputation-with-scikit-learn-8f3eb22b1a38

ExtractTree Inputation (Categorical Variables)

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html#sklearn.ensemble.ExtraTreesRegressor

https://towardsdatascience.com/going-beyond-the-simpleimputer-for-missing-data-imputation-dd8ba168d505