Health Analytics Challenge Lab: NHANES

Precision Analysis for Chronic Diseases

IDEA

Rensselaer Institute for Data Exploration and Applications

Authors/Group Members:

Jesse Rotering, Mason Armstrong, Sankalp Chauhan (Sonny) , and Fitz (Yuan Gao)

Rensselaer Polytechnic Institute

Abstract:

The concept of precision health is a very promising field. Using the information from National Health and Examination Survey (NHANES), certain variables were extracted with respect to physical activity and nutritional data, and then this data was combined with data related to the chronic diseases: systolic blood pressure, diastolic blood pressure, coronary heart disease, breast cancer, and type 2 diabetes. This data was ingested into the risk analyzer and an analysis was made regarding the risk factors present for certain diseases with respect to bmi groups, gender, and ethnicity for the general population.

Introduction/Background:

There are a number of chronic diseases that plague our society but it's often hard to track the factors or maybe the specific cause for an individual who developed it. For our work we focused on data regarding nutritional intake and physical activity from the National Health and Nutrition Examination Survey (NHANES). NHANES takes surveys every year to gather information on a number of subjects to get a good view of how the population looks with respect to health and lifestyles.

Problems Tackled:

The goal of this project is to develop a tool to perform dynamic precision analysis of chronic diseases and find risk factors for subpopulations and the population as a whole.

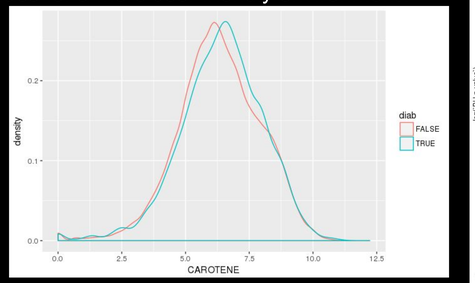
Questions addressed:

- What physical activity/nutritional variables are associated with chronic diseases?
- How can we update the app interface for a more user friendly environment?

Data Description:

NHANES data

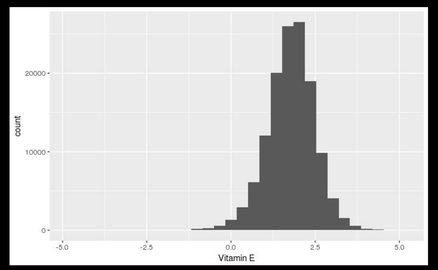
- Categorical and continuous values
- Taken on the entire population
- Stratified data collection
- Information on nutritional intake as well as physical activity
- added 16 new physical activity variables and 17 new nutritional variables



This is a density plot of carotene intake with diabetics and non diabetics. this displays a significant difference that can be used to determine if someone may be or will become diabetic

Data Cleaning

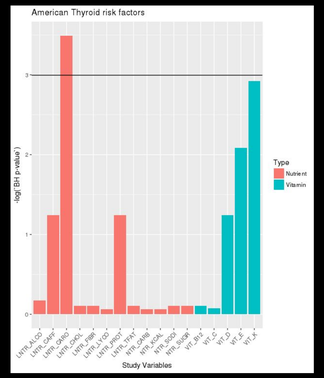
- Account for missing values
- Stitch together similar variables
- Transform skewed data
- Remove unnecessary variables
- Adjusted survey weights to account for multiple years of surveys put together



a histogram showing the distribution of vitamin E intake among the population. it is fairly normal so no transform is needed and it can be packaged up for the analysis by the risk analyzer.

Data Analytics Methods:

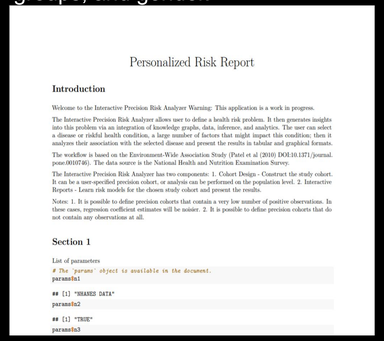
First the physical data was cleaned, by removing all columns that had more than 80% missingness. Afterwards the physical activity data was broken down into three groups using the missingness heatmap that was given in dataconstruction.Rmd. Categorical variables with low answers were removed, since they had very high log odds ratio and only represented a small portion of the population. A data dictionary as well as table was created in order to differntiate between categorical and numerivcal variables in the risk analyzer. With the nutritional data, after checking for important nutritional variables from the NHANES documentation, logistic regression was used to determine important factors . Afterwards the nutritional data were grouped by similar missingness using the documentation on the NHANES website.



This is a bar graph shows relative BH p-values of diffent nutritional factors and their relation to thyroid condition on the population as whole. plots like these are made in the risk analyzer and you can see what factors have the highest association and a line is given on the graph to help evaluate what is considered highly associated. These plots are then made for other subpopulations that will be analyzed in the same way and comparisons are made that we can draw higher level conclusions about with regards to how different groups of people are affected with chronic illnesses.

Experimental Results:

The risk analyzer is used to find risk factors for diseases thyroid condition, type 2 diabetes, systolic blood pressure, diastolic blood pressure, breast cancer, and coronary heart disease given control variables, such as age, income, ethnicity, BMI, gender, serum cotinine, and menopausal status. One of which was based on certain subpopulations. In order to see what variables are important we made use of the *p*-values, and set the threshold to 0.05. Afterwards the risk factors were shown for the certain diseases with respect to the whole population. The results of what was found is shown in the table below. Later we found risk factors for certain diseases with respect to subpopulations like bmi groups, ethnicity groups, age groups, and gender.



Personalized Risk Report

Introduction

Welcome to the Interactive Precision Risk Analyzer Warning: This application is a work in progress.

The Interactive Precision Risk Analyzer allows user to define a health risk problem. It then generates insights into this problem via an integration of knowledge graphs, data, inference, and analytics. The user can select a disease or riskful health condition, a large number of factors that might impact this condition; then it analyzes their association with the selected disease and present the results in tabular and graphical formats.

The workflow is based on the Environment-Wide Association Study (Patel et al (2010) DOI:10.1371/journal.pone.0010746). The data source is the National Health and Nutrition Examination Survey.

The Interactive Precision Risk Analyzer has two components: 1. Cohort Design - Construct the study cohort. It can be a user-specified precision cohort, or analysis can be performed on the population level. 2. Interactive Reports - Learn risk models for the chosen study cohort and present the results.

Notes: 1. It is possible to define precision cohorts that contain a very low number of positive observations. In these cases, regression coefficient estimates will be noisier. 2. It is possible to define precision cohorts that do not contain any observations at all.

Section 1

List of parameters

```
# The `params` object is available in the document.
params$s1

## [1] "NHANES DATA"
params$s2

## [1] "TRUE"
params$s3
```

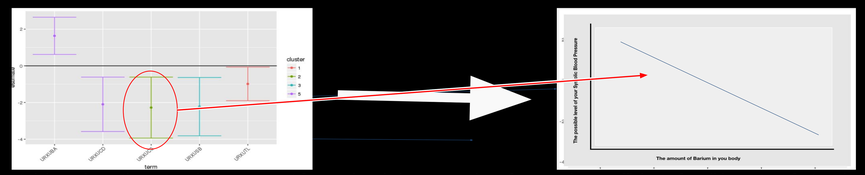
Personal risk report

Type 2 diabetes	Breast Cancer	Coronary Heart Disease	Thyroid Condition	Systolic Blood Pressure	Diastolic Blood Pressure
Cholesterol	None	less active than most people your age	about the same active past 30 days than in past 12 months	Cholesterol	Cholesterol
Protein		Unable to do tasks around home/yard for at least 10 minutes in the last 30 days		Alcohol	Protein
Total Sugars		Unable to do muscle strengthening activities for at least 10 minutes over the past 30 days.		Energy (kcal)	Alcohol
Less active than most people your age		unable to do moderately intense activities for at least 10 minutes over past 30 days			Caffeine
Do not engage in vigorous intensity sports for at least 10 minutes on a daily basis		unable to do vigorous activities over the past 30 days for at least 10 minutes.			Total Fat
		do not walk or use bicycle for at least 10 minutes to get to places			Carbohydrate
		Hours watching tv and computer use in past 30 days			Sugars
					Energy(kcal)

These are some examples of risk factors for chronic illnesses for the population as a whole

Discussion:

The findings of the risk analyzer and secondary analysis of those outputs can be important for determining how a sub population is affected by certain lifestyle factors causing chronic illness. The tests can be run on a wide array of factors and the personalized output method can make the response important for a common person or even a healthcare professional.



With the current risk analyzer the output is more expert focused and shows a series of regression coefficients for all important factors but for a normal patient a new plot should be made that can easily show the effect of a single factor.

Conclusion & Recommendations:

The whole point of the shiny app that analyzes the data is to make recommendations to provide context and suggestions on what changes can be made to improve the life of an individual or maybe even a larger community as a whole. In the end expanding the functionality of this app will help it be a tool to help people in all sorts of groups.

References/Citations:

Alexander, Breneman, Curt, Bennett, and Kristin P. "Cadre Modeling: Simultaneously Discovering Subpopulations and Predictive Models." [1402.1128] Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. February 07, 2018. Accessed August 08, 2018. https://arxiv.org/abs/1802.02500.

Hall, M. A., S. M. Dudek, R. Goodloe, D. C. Crawford, S. A. Pendergrass, P. Peissig, M. Brilliant, C. A. McCarty, and M. D. Ritchie. "Environment-wide Association Study (EWAS) for Type 2 Diabetes in the Marshfield Personalized Medicine Research Project Biobank." Advances in Pediatrics. Accessed August 08, 2018. https://www.ncbi.nlm.nih.gov/pubmed/24297547.

"National Center for Health Statistics." Centers for Disease Control and Prevention. July 31, 2018. Accessed August 08, 2018. https://www.cdc.gov/nchs/nhanes/index.htm.

Ng, K., J. Sun, J. Hu, and F. Wang. "Personalized Predictive Modeling and Risk Factor Identification Using Patient Similarity." Advances in Pediatrics. March 25, 2015. Accessed August 08, 2018. https://www.ncbi.nlm.nih.gov/pubmed/26306255.

Patel, C. J., J. Bhattacharya, and A. J. Butte. "An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus." Advances in Pediatrics. May 20, 2010. Accessed August 08, 2018. https://www.ncbi.nlm.nih.gov/pubmed/20505766.

Welcome to the Interactive Precision Risk Analyzer

Warning: This application is a work in progress.

The Interactive Precision Risk Analyzer allows user to define a health risk problem. It then generates insights into this problem via an integration of knowledge graphs, data, inference, and analytics. The user can select a disease or riskful health condition, a large number of factors that might impact this condition; then it analyzes their association with the selected disease and present the results in tabular and graphical formats.

The workflow is based on the Environment-Wide Association Study (Patel et al (2010) DOI:10.1371/journal.pone.0010746). The data source is the National Health and Nutrition Examination Survey.

The Interactive Precision Risk Analyzer has two components:

1. Cohort Design - Construct the study cohort. It can be a user-specified precision cohort, or analysis can be performed on the population level.

2. Interactive Reports - Learn risk models for the chosen study cohort and present the results.

Notes:

1. It is possible to define precision cohorts that contain a very low number of positive observations. In these cases, regression coefficient estimates will be noisier.

2. It is possible to define precision cohorts that do not contain any observations at all.

Interactive Precision Risk Analyzer

Find Known Cohort Risks

Perform precision cohort analysis?

No

Age Category

Younger Adults | Older Adults | Elderly

Gender

Female

Ethnicity

Mexican American | Other Hispanic | Non-Hispanic White | Non-Hispanic Black | Other

BMI Category

Healthy weight | Underweight

Survey Cohort Years

1999-2000 | 2001-2002 | 2003-2004 | 2005-2006 | 2007-2008 | 2009-2010

Diseases:

Diabetes-2

Data Sources:

NHANES DATA

Use control variables from knowledge graph?

Yes

Query knowledge graph for other sets

If not, select control variables:

Age BMI

Environmental Factors:

Dietary Factors:

ntr ntrm vit

Physical Activity Variables

Set 1 Set 2 Set 3

Search TOXNET

p-value threshold for significance

Analysis summary for diabetes-2 :

Number of study variables: 31
Control variables: age, gender, ethnicity, income, BMI, urinary creatinine (when study variable is urine-based)

Study cohort summary:

Genders: female, male
Age categories: children, younger adults, older adults, elderly
Ethnicities: Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Other
BMI categories: underweight, healthy weight, overweight, obese
Survey years: 1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013

Compare results?

The risk factors for diabetes-2 are:

Show 25 entries Search:

Study Variable Code	Regression Coefficient	Std Error	nPos	nObs	BH p-value	Study Variable	Category	component
PAQ520v2	0.447338778	0.14931257	775	9117	1.492074e-02	(MEC Interview Version) Compared with most {boys/girls} {your/SP's} age, would you say that {you are/SP is}... (SP Interview Version) Compared with most {men/boys/women/girls} {your/SP's} age, would you say that {you are/s/he is} . . .	Physical Activity	Questionnaire
PAQ520v3	0.187369675	0.12872938	775	9117	3.398357e-01	(MEC Interview Version) Compared	Physical Activity	Questionnaire

Personalized Risk Report

Introduction

Welcome to the Interactive Precision Risk Analyzer Warning: This application is a work in progress.

The Interactive Precision Risk Analyzer allows user to define a health risk problem. It then generates insights into this problem via an integration of knowledge graphs, data, inference, and analytics. The user can select a disease or riskful health condition, a large number of factors that might impact this condition; then it analyzes their association with the selected disease and present the results in tabular and graphical formats.

The workflow is based on the Environment-Wide Association Study (Patel et al (2010) DOI:10.1371/journal.pone.0010746). The data source is the National Health and Nutrition Examination Survey.

The Interactive Precision Risk Analyzer has two components: 1. Cohort Design - Construct the study cohort. It can be a user-specified precision cohort, or analysis can be performed on the population level. 2. Interactive Reports - Learn risk models for the chosen study cohort and present the results.

Notes: 1. It is possible to define precision cohorts that contain a very low number of positive observations. In these cases, regression coefficient estimates will be noisier. 2. It is possible to define precision cohorts that do not contain any observations at all.

Section 1

List of parameters

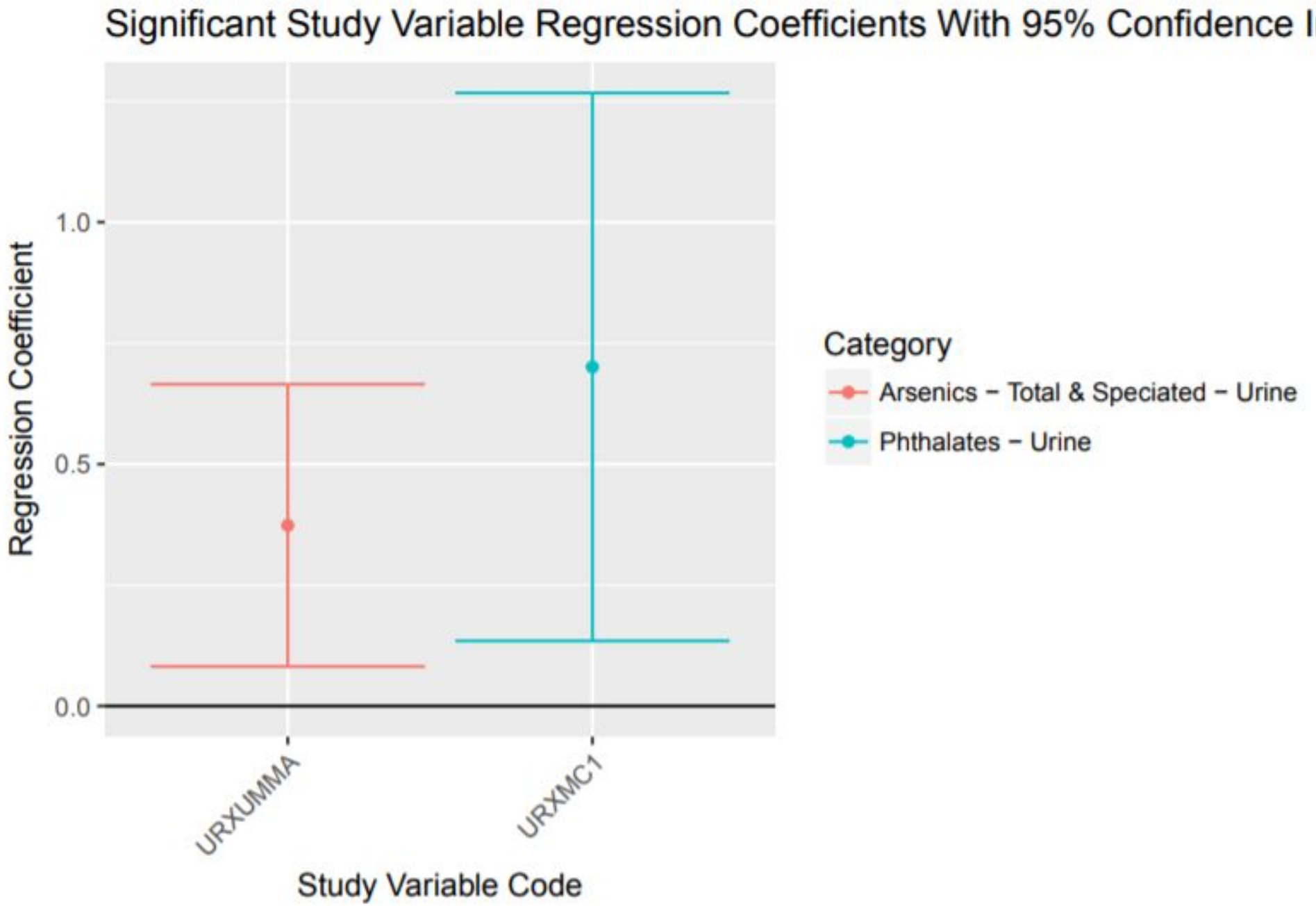
```
# The `params` object is available in the document.
params$n1
```

```
## [1] "NHANES DATA"
```

```
params$n2
```

```
## [1] "TRUE"
```

```
params$n3
```


Significant Study Variable Regression Coefficients With 95% Confidence In
Regression Coefficient
1.0
0.5
0.0
URXUMMA
URXMC1
Study Variable Code
Category
Arsenics – Total & Speciated – Urine
Phthalates – Urine

Secure | https://lp01.idea.rpi.edu/rstudio-ose/?view=shiny

tps://lp01.idea.rpi.edu/rstudio-ose/p/7454/ | Open in Browser | Publish

Interactive Precision Risk Analyzer | Introduction | Personal Information | For Patient | For Professionals

Basic Information

Name

Bob

Birth year

1993

Today's year

2018

BMI

Gender: Female

Ethnicity: Non-Hispanic White

Height

5 ft 3 in

Weight

150 lb

Annual Income

$ 1.5

Control factor

Urinary Creatinine

151

Risk factor

SDMVPSU

2

SDMVSTRA

18

LBXCOT

0.652

sbp

110.6668

dbp

65.33333

URXUTU

0.03

URXUTL

0.21

URXUSB

0.029

URXUPB

0.4

URXUMO

Diseases you want to learn about (Potential Disease)

Diseases you may have:

Systolic blood pressure

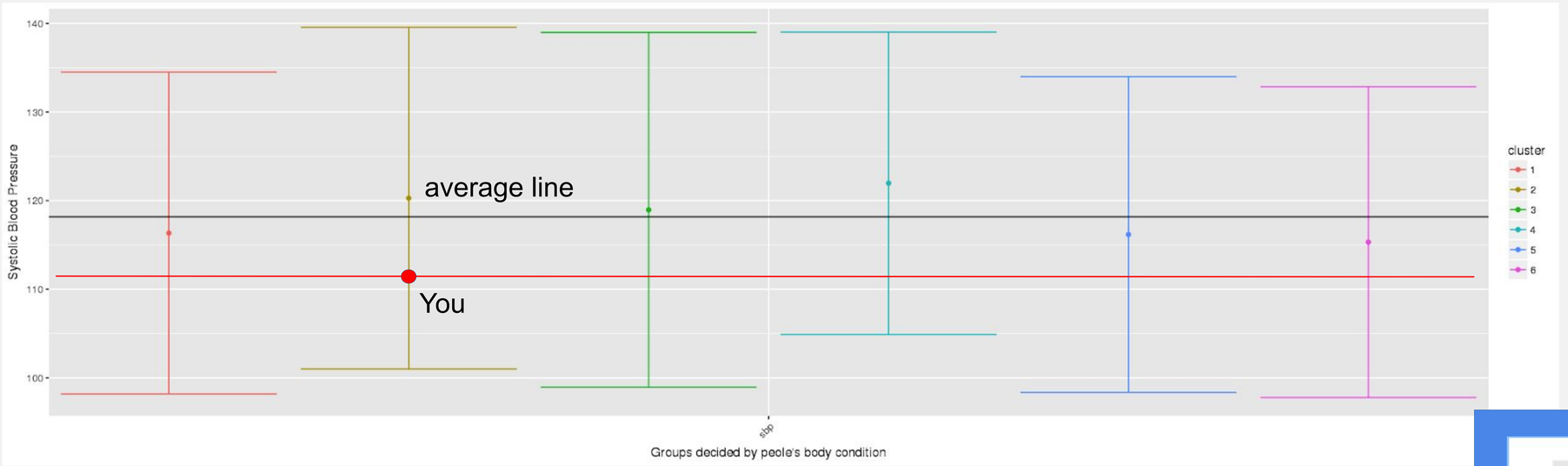
Apply Changes

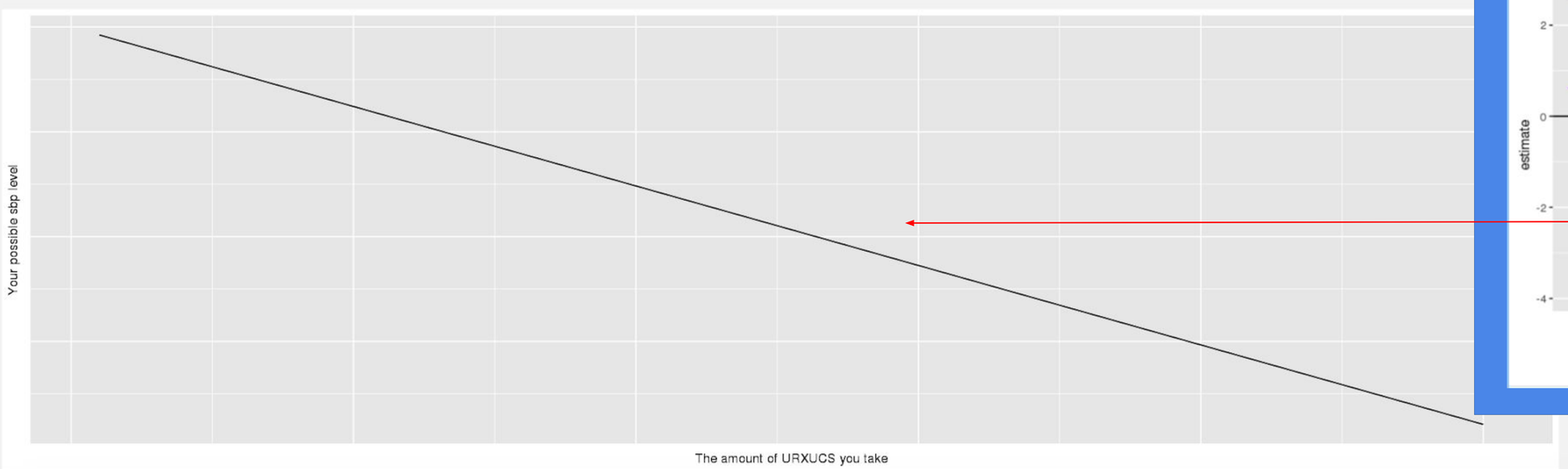
Personalized Input Page

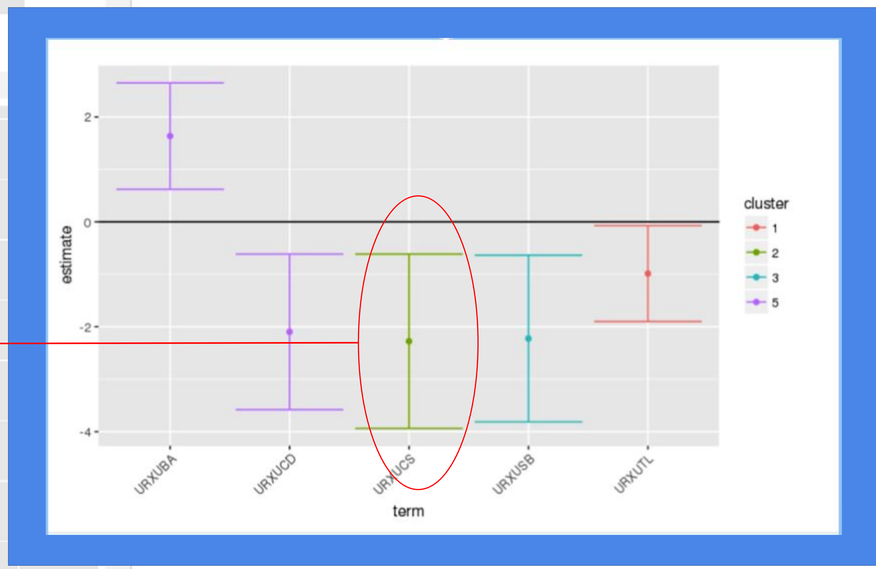
Sublime Text

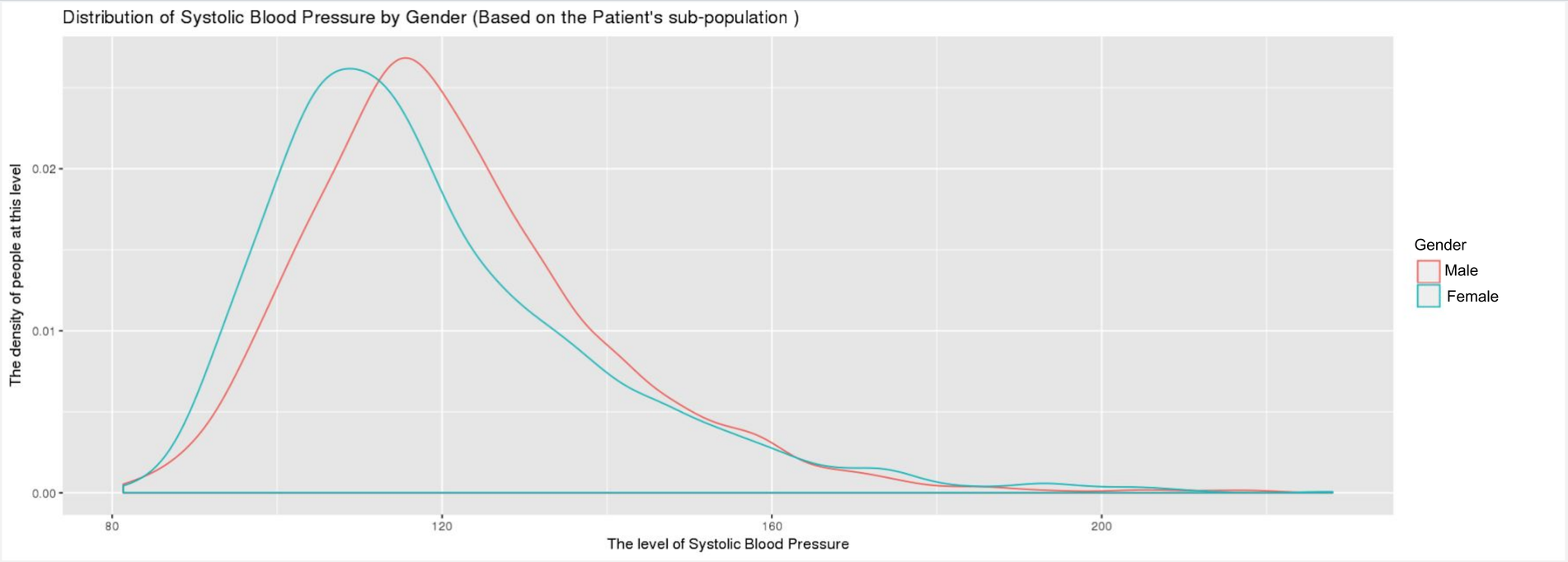
This is the Health Condition Report for Bob

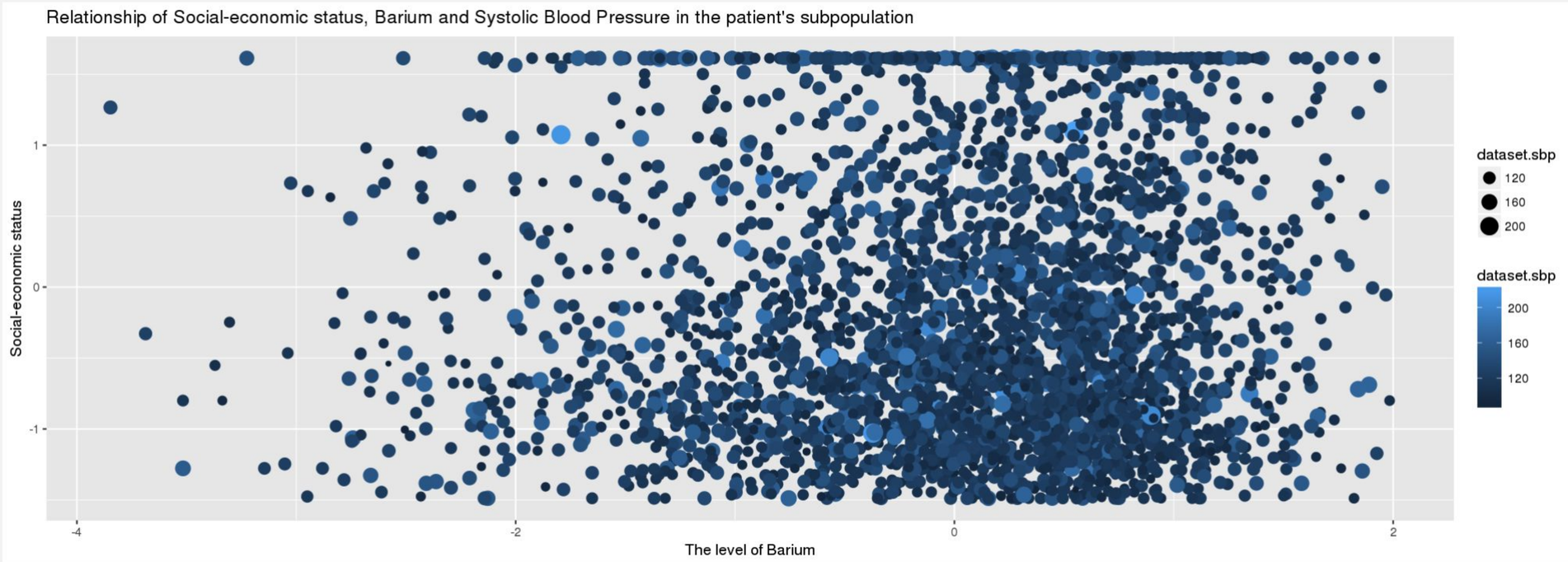
You belong to group 2

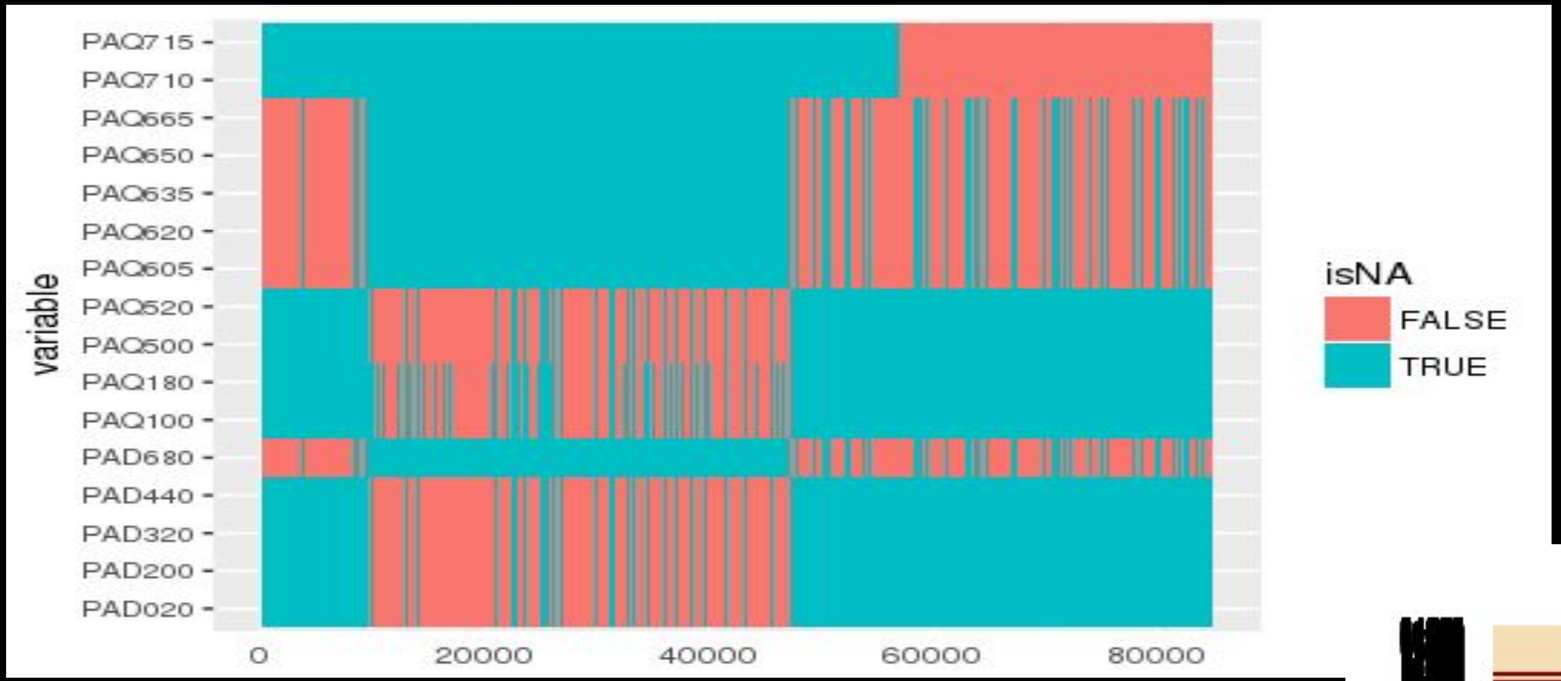


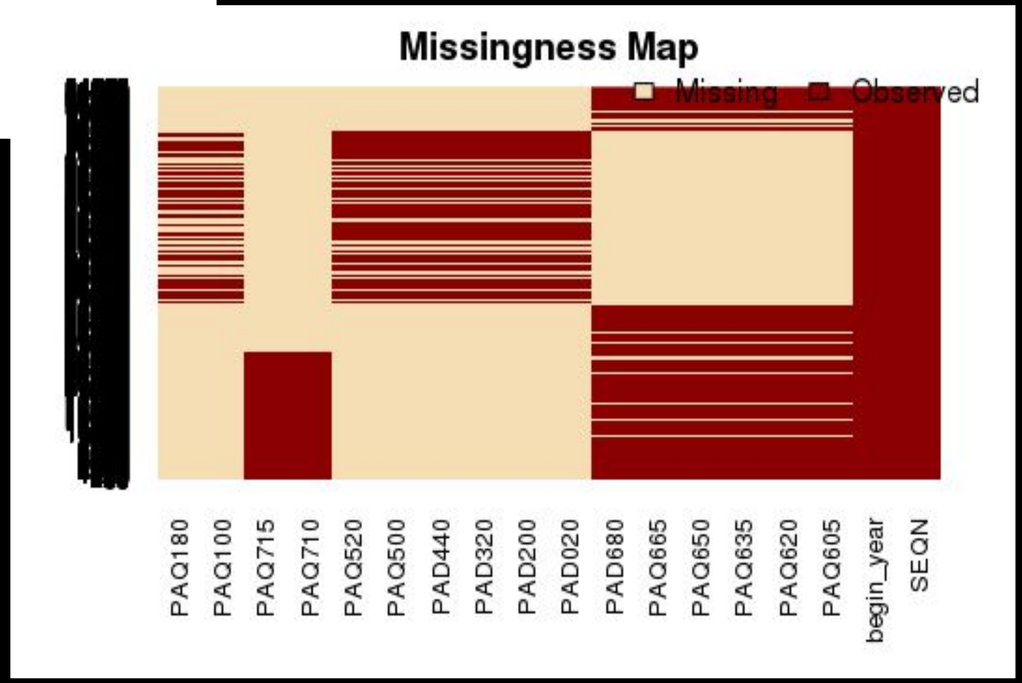





Distribution of Systolic Blood Pressure by Gender (Based on the Patient's sub-population)
The density of people at this level
0.02
0.01
0.00
80
120
160
200
The level of Systolic Blood Pressure
Gender
Male
Female


Relationship of Social-economic status, Barium and Systolic Blood Pressure in the patient's subpopulation
Social-economic status
1
0
-1
-4
-2
0
2
The level of Barium
dataset.sbp
120
160
200
dataset.sbp
200
160
120


PAQ715
PAQ710
PAQ665
PAQ650
PAQ635
PAQ620
PAQ605
PAQ520
PAQ500
PAQ180
PAQ100
PAD680
PAD440
PAD320
PAD200
PAD020
variable
0
20000
40000
60000
80000
isNA
FALSE
TRUE


Missingness Map
Missing
Observed
PAQ180
PAQ100
PAQ715
PAQ710
PAQ520
PAQ500
PAD440
PAD320
PAD200
PAD020
PAD680
PAQ665
PAQ650
PAQ635
PAQ620
PAQ605
begin_year
SEQN