THE BRDIGE

DATA SCIENCE

DNA Analysis & health. Exploratory Data Analysis (EDA)

Erika Kvalem Soto

April 28, 2021

Abstract

The purpose of this work is to perform an exploratory data analysis (EDA). This approach analyzes the data to summarize the main characteristics using different visualization and statistical methods to get an idea of what type of valuable information the dataset contains.

In this case the dataset that will be explored are two genome data sets corresponding to two different individuals. The idea is to be able to understand their genome in order to predict the probability of developing one of the top ten deadly diseases worldwide.

1 Why?

In 2019, the top 10 causes of death accounted for 30.7 million deaths worldwide. The top global causes of death, in order of total number of lives lost, are associated with cardiovascular and coronary heart diseases, respiratory chronic diseases and neonatal conditions [1]. Causes of death can be grouped into communicable(infectious and parasitic diseases), noncommunicable (chronic) and injuries.These diseases are linked to our genome and their likelihood can be studied through a DNA analysis [1].

Leading causes of death globally 2019 [1]:

- Cardiovascular diseases (CVDs)
- Coronary heart diseases (CHDs)
- Chronic obstrucitve pulmonary disease
- Lower respiratory infections
- Neonatal conditions
- Trachea, bronchus, lung cancers
- Alzheimer's disease and other dementias
- Diarrhoeal diseases
- Diabetes mellitus
- Kidney diseases
- Stroke

1.1 Business case: Prevention avoids cost

Cardiovascular and Coronary heart diseases: As explained above, these are the top deadliest diseases in the world and they entail a huge cost to the healthcare system. In this section I try to bring light on how much these diseases currently cost and how this figure could be lowered by using prevention [2] .

For the case of United States, the average hospital charge for a heart operation or related procedure is about $85,000. Concerning this issue, a study was published in the Feb. 1, 2017, Journal of the American Heart Association focused on Medicare costs[3]. The study authors estimated that If all Medicare beneficiaries followed some of the heart-healthy habits to reduce cardiovascular disease, it would save more than $41 billion a year in Medicare costs.

Chronic diseases: It is clear that preventive services at the primary and secondary levels yields results in net medical savings to the healthcare system. It is also important to prevent beyond costs alone to include value and benefits perceived by quality of life of the patients [4].

For example, primary preventive services, such as daily aspirin use and alcohol and tobacco use screenings, yielded net savings of nearly 1.5 billion dollars. Eventhough these services carry a cost they have a certain positive impact on health [2].

Individuals with one or more chronic conditions account for approximately $1.5 trillion in healthcare spending per year [4]. Focusing on high-risk patients with chronic conditions offers high savings and cost-effectiveness margins because the likelihood of needing high-cost treatments are far greater than the costs incurred by provision of preventive services [3]. These services could produce substantial savings, perhaps as much as 45 billion per year, having a strong potential for improving health and reducing spending.

Conclusion: These are just two examples of the most deadly type of diseases among the top ten which kill 30.7 million people in 2019 [1]. In data on the rest was exposed it could be inferred that prevention is a big asset on fighting against these diseases. Moreover, thanks to the technology available nowadays and the latest state-of-the-art findings on genetics prevention can be even cheaper and easier. It is our responsibility to make it visible and use it wisely for the betterment of society.

2 How?

23andMe and MyHeritage are two among many private companies, in this case USA

based, that are devoted to the analysis of DNA sequences to generate reports relating to the customer's ancestry and genetic predispositions to health-related topics [5]. They provide a genetic testing service in which customers provide a saliva sample that is laboratory analysed, using single nucleotide polymorphism (SNP) genotyping [6].

2.1 Raw data

For the SNP genotype analysis provided by these two companies I have been able to work with two different data set corresponding to Zeeshan-ul-hassan Usmani's (23andme) and Daniel Gago Castro (MyHeritage). The first I found on the Kaggle web page publicly posted and the second I asked my classmate Daniel Gago Castro who have previously performed this analysis for his use consent.The data is composed of aprroximately 600000 SNPs and it gives information about the location, the gene and the genotype (the nucleotides associated)

3 Final conclusion

For both individuals I have explored their genome [7] from easiest phenotype features to more complex and I have studied the SNPs coding for the top 10 most deadly diseases as mentioned above [1]. These have been my observations according the matching codes proposed by SNPedia:

SNP	Feature studied	Zeeshan Phenotype	Daniel Phenotype
rs1800401	Eye color	green/hazel/brown/black eyes	-
rs12913832	Eye color	-	brown eye color
rs1426654	Skin color	probably light skin	probably light skin
rs16891982	Skin color	7x black hair	7x black hair
rs1333049	CVDs	1.5x risk CVDs	1.5x risk CVDs
rs12425791	Stroke	normal profile	normal profile
rs3736309	Chronic pulmonary disease	0.44X decreased risk	normal profile
-	Lower respiratory infections	too complex to include	too complex to include
rs1800629	Neonatal conditions	generally normal risk	higher risk
rs8034191	Lung cancer	normal profile	1.27x risk
rs145999145	Alzheimer's	normal profile	normal profile
rs1470579	Diabetes Type II	1.2x increased risk	1.2x increased risk
rs17319721	Kidney diseases	normal profile	normal profile

References

[1] *The top 10 causes of death*. eng. 2020.

[2] *The value of prevention*. eng. 2017.

[3] *Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association*. eng. 2017.

[4] *Priorities among effective clinical preventive services: results of a systematic review and analysis*. eng. 2017.

[5] *Pharmacogenetic testing through the direct-to-consumer genetic testing company 23andMe*. eng. 2017.

[6] *The essence of SNPs*. eng. 199.

[7] *I have Had My DNA Tested. . . Now What?* eng. 2018.