CA' FOSCARI UNIVERSITY OF VENICE
DEPARTMENT OF ECONOMICS

Master's Degree in Data Analytics for Business and Society

Unleashing the Power of Big Data: Revolutionizing Healthcare for a Brighter Future

Elisabetta Romano

Supervisor: Stefano Campostrini
Co-supervisor: Angela Andreella

Academic year 2022-2023

ACKNOWLEDGEMENTS

Writing the dissertation is always a hard and necessary step that leads to questioning yourself and actually figuring out what notions you have fully understood.
I hope that this thesis of mine will be able to bring me to a dignified conclusion to this journey. I thank all the people who have been close to me in this journey and especially those who have always managed to support me in difficult times.
I would like to thank the distinguished professors Angela Andreella and Stefano Campostrini for their availability and great help.
I would also like to thank all the people who have accompanied me on this journey, both new and old.
Choosing to start all over again is never make it easy, especially away from home. However, I thank the fate that made me meet people who have always supported and helped me, especially I want to thank Claudia, Constanza, Gabriele and Elena.
Without whom, my journey would have been empty and far more difficult.
I also want to thank Rita and Jessica, Abi and Enrica, who despite the distance have always managed to be there for me.
But in particular I would like to thank Hali for being my best roommate, for the evenings on the balcony, and for always giving me time to listen to my ramblings.
My stop in Venice was a carousel of emotions but I do not regret a single day of this experience, because it it made me the best version of me.

ABSTRACT

The thesis will investigate the impact of big data on health policy, by discussing and proving that collecting and analyzing data is helpful in improving the health population. Health Data Science, an ensemble of statistics, econometrics, and health policy, is the subject used to do this research. In particular, there will be a focus on health surveys and data surveillance, two important tools used to make inquiries about public health. One of the main aspects studied during the health survey is risk factors, behavioral actions that affect the health of the individual such as smoke, alcohol, and an improper diet. To support the thesis the Behavioural Risk Factor Surveillance System, an American survey based on a population of 200,000 citizens, with questions about general health, demographics, screening, and risk factors, will be analyzed. After cleaning, preprocessing, and exploring the dataset, spatial clustering is applied to find groups of USA states sharing a common socio-demographic and health structure. The `skater` algorithm is applied, which is a spatial clustering approach that uses the graph theory to take into account the spatial structure of the data in creating the groups of observation. The results of the spatial cluster analysis will be interpreted by observing the distribution of demographic factors, risk factors, and cancer among the clusters. To conclude, the last chapter will discuss the challenges and possible directions of health surveys and, in detail, of BRFSS survey.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

Page

ABSTRACT ii

LIST OF TABLES iii

LIST OF FIGURES v

CHAPTER 1: Impact of Big Data in Health Data Science 1

1.1 What Is Health Data Science? 1

1.2 Big Data 3

1.3 Risk Social Determinant 5

1.3.1 California Endowment 6

1.3.2 Health Care Surveys and Data Surveillance 7

1.4 Administrative Data and Use of Big Data 9

1.4.1 The IHI Triple Aim initiative 10

1.4.2 Machine Learning's Effect on Healthcare Policy: Lessons from Five Case Studies 11

1.4.3 Data Protection in the USA 12

CHAPTER 2: From Data to Action: Using BRFSS to Understand Risk Factors and Promote Health 13

2.1 What is Disease Prevention? 13

2.1.1 Global Burden of Diseases 16

2.2 Health Care Policy in the USA 18

2.3 BRFSS: America's Leading Health Survey System 19

2.3.1 Assessing the Reliability of BRFSS 21

2.3.2 Safeguarding Data Privacy in the BRFSS 22

CHAPTER 3: Behavioral Risk Factor Surveillance System Data Analysis . . 24

3.1 Research Question 24

3.1.1 Data 25

3.2 Exploring the Depths: BRFSS Data through Explanatory Data Analysis 26

3.2.1 Demographic Variables 30

3.2.2 Behavioural Risk Factor Analysis 35

3.2.3 Cancer Disease Analysis 40

3.3 Spatial Clustering 41

3.3.1 Graph theory, Minimum Spanning Tree and Skater algorithm . . 42

3.3.2 R Code Analysis 44

CHAPTER 4: Unveiling the Path to Healthier Communities: The Future of BRFSS and Health Surveys. 49

4.1 Values, Challenges, and Future Directions of Health Survey, Data Surveillance 49

4.2 Exploring the Use of Big Data in Healthcare Analysis 50

4.3 Spatial Clustering and Personal Observation 52

LIST OF TABLES

Table	Title	Page
3.1	List of dataset variables, unique level for each variable, and the top counts of the two values with the higher frequency.	28
3.2	Absolute frequency distribution for state and sex.	28

LIST OF FIGURES

Figure	Title	Page
1.1	Vulnerability and unmet healthcare needs: the influence of multiple risk factors. Source: Big Data and Health Analytics, by Marconi and Lehmann [1].	5
1.2	Triangle of the Triple Aim, image taken from [2].	10
3.1	Age distribution of the respondents divided into six levels, in absolute frequency.	30
3.2	Old adult distribution for each state in relative frequency.	31
3.3	Income distribution for age group, conditional relative frequencies per age	32
3.4	Education level distribution for each age, in absolute frequency, where 1 is Never attended school, 2 is elementary, 3 is attending high school, 4 is High school graduate, 5 is college or technical school, and 6 is for college graduate.	32
3.5	Eight types of employ, in relative frequency.	33
3.6	Employ distribution per age and sex, where in the x-axis 1 stands for employed, level 2 for self-employed, level 3 for out of work from more than one year, level 4 for out of work from less than one year, level 5 for homemaker, level 6 for student, level 7 for Retired, and level 8 for unable to work.	34
3.7	Health coverage ownership in absolute frequency for each age group, where 1 stands for yes and 2 stands for no.	34
3.8	Health coverage distribution for each state in relative frequency.	35
3.9	Risk factor number 1: Alcohol use, where level 1 refers to not drinkers respondents and level 2 for drinkers, distribution of drunker per age and sex, absolute frequency.	36
3.10	Risk factor number 2: BMI relative frequency, where the percentage of underweight respondents is about 1% and overweight and obese respondents is about 70%.	37
3.11	Risk factor number 3: smoke use where level 1 refers to smokers and level 2 to respondents who don't smoke, in absolute frequency.	37
3.12	Risk factor number 4: Physical activity, where level 1 refers to respondents who practice regular physical activity and level 2 refers to those who do not, distribution per age and sex in absolute frequency.	38

3.13 Risk factor distribution for each state, the x-axis values are relative frequency. 39

3.14 Risk distribution per age. 39

3.15 Cancer distribution relative frequency per age and sex, where 1 ion x-axis means male and 2 means female, in the legend 1 means the observation have or had cancer and 2 means not. 40

3.16 Cancer distribution among the state, in relative frequency. 41

3.17 USA initial graph with the edge and the centroids, used to do the spatial clustering analysis. 45

3.18 USA MST graph representation. 46

3.19 Silhouette plot where the y-axis is the silhouette score, the x-axis is the number of cuts considered, i.e., the times the dataset is cut, where considering 2 cuts means dividing the dataset into 3 clusters. 46

3.20 The plot shows the result of the clustering analyses, where the three different colors refer to three relative clusters.). 47

3.21 Box-plot of 11 scaled variables for each cluster defined in Figure 3.20. . 48

CHAPTER 1

Impact of Big Data in Health Data Science

Thanks to the advent of big data, the phrase "to have information is to have power" has taken on new meaning: through its use, it is indeed possible to regain new insights and affix changes previously thought unattainable. With the right information, you can choose the best strategy to reach your goal. Information nowadays is essential to do everything, and for this reason, big data is rapidly increasing its importance in several areas. One of the areas in which they are essential is healthcare, and this will be the main argument of this thesis. Big Data has become a buzzword, with many businesses and startups touting the benefits of mining massive amounts of information. However, the healthcare sector has seen tangible benefits from big data, where the principal aim is to reduce hospital and health centers' costs, increase the quality of patient care, and use prevention to improve public health. This chapter will explain the role of health data science, how the use of data coming from health surveillance systems helps in the prevention of diseases, and the impact of risk social determinants in public health care.

1.1 What Is Health Data Science?

The combination of mathematical, computer, statistical, and epidemiological subjects gives rise to health data science, a subject of study that aims to improve scientific

knowledge in health care and to analyze and study all those quantitative skills that apply in the analysis of health data. The main goal is to investigate medical and health subjects through the use of digital methods and principles, such as data collection and analysis, the use of statistical models and artificial intelligence, and the use of analytical techniques to improve surgical operations. Since the collection and integration of data originate from multiple sources, one of the goals is to create a common dictionary that allows communication from data with different origins, such as regional healthcare hospitals and scientific societies. There are two main purposes of health data science: the first is the collection and analysis of hospital data that focus on the individual and are collected by physicians and researchers, while the second research question focuses on the public, drawing information from questionnaires filled out by a large part of the population. Both will be analyzed in this section, but the entire thesis focuses exclusively on data collected from and for the population by proper questionnaires.

The administrative data regard individual healthcare and are collected in hospitals, in physician's offices, or in community pharmacies. They are especially used for administrative aims but also to investigate the cost and the benefit of healthcare delivery, to improve services, and to increase the overall effectiveness of healthcare facilities. This kind of data is very efficient because they are collected and updated daily, a quality that makes them easy to find. Through data analysis, healthcare costs can be reduced in a number of ways. An example might be comparing and analyzing different methods of treatment for a disease from which to derive the most efficient and least costly treatment for the patient. Alternatively, through analyses to evaluate the prices offered by incumbent suppliers while trying to achieve higher savings margins. To support the collection of administrative data, in 2010, was enacted a law by President Barack Obama : the "Patient Protection and Accountable Care Act (PPACA)" or "Obamacare'. This law provides numerous rights and protections that make health coverage more equitable and affordable, and it also includes a statement to improve the quality of care, reduce healthcare costs, and promote prevention and wellness through the collection of big data, the new source of information used to make decisions. If it is fully implemented, the number of uninsured Americans is expected to be reduced by more than half.

The huge amount of data is also useful and crucial in public health surveillance, as it contributes to control and sudden response in case of disease swarms and epidemics. The use of big data in healthcare can lead to greater cost savings and better resource management. With an in-depth understanding of patient populations and their specific needs, clinicians can improve workflows, streamline operations, and allocate resources more efficiently. Thanks to big data, the final customer becomes active in the role of self-care for their own health, monitoring diet, medication adherence, and other details in order to follow evidence-based care that prioritizes the individual, allowing for more effective

action. There are so many in-house trying to collect as much data as possible and find the most effective ways to get all the information. Among them is Intermountain Healthcare [3], a not-for-profit health system with the aim to provide high-quality and low-cost health care that collects information from as many as 22 hospitals, including cost, time, and disease management methods. Personal data has emerged as a new source of economic value as it provides significant information for companies about individuals' interests and activities, which is valuable for targeted advertising. The collection, aggregation, analysis, and monetization of personal data have become crucial in this area. While generating profit from personal data might seem invasive to some, there is no legal reason to prevent a business model based on processing personal information from developing. Nevertheless, there are very strict privacy limitations, which can restrict the revenue of such business models. The question that remains is what privacy laws protect and forbid others from doing when it comes to online companies that can easily gather data from their users for profit-making advertising. The ultimate goal of collecting data about the public is to fight against inadequate care and poor quality of care. These failures are common globally in different conditions and countries, but they hit the frailest part of the population the hardest. Most public data are collected through questionnaires and data surveillance.

Systematic deficits exist in the quality of care, with major gaps in safety, prevention, integration, and continuity. Those who suffer from poor and, often, absent health care are the most fragile groups, including prisoners, the poor, the uneducated, and those with stigmatized conditions. Health systems must prioritize measuring and reporting on what matters most to people: skilled care, user experience, health outcomes, and trust in the system. Appropriate measures of health system quality must be generated and used at the national and sub-national levels, and performance must be reported annually to the public through a dashboard of key metrics. Therefore, reliable data systems and institutions with strong quantitative and analytical capabilities are fundamental to solving and addressing these problems.

1.2 Big Data

But what is big data, and how can we create it? Many are the definition of big data, but generally, as George Seffers, the Magazine Editor in Chief that writes several articles about the progress of big data, says in "The global race of big data": they simply are "data sets too large for typical database software tools to capture, store, manage, and analyze" [1]. In other words, big data is the result of the action of billions of sources from which a computer, GPS device, cell phone, and medical devices can take information about the analyzed topic.

Big data analyses are very important in today's world due to several reasons:

- Directing strategic decisions: big data analytics constantly analyzes huge volumes of varied data at high speed, allowing companies to derive valuable insights. Thanks to this, any business can make its decisions from a very informed base and reach its intended strategy as effectively as possible.

- Better knowledge of their customers: the use of big data allows a company to learn more about its customers through the use of the data they provide to them. This data makes it possible to create appropriate customer profiling, which facilitates the identification of preferences, buying patterns, and behaviors. As a result, marketing campaigns are personalized, and their use is far more efficient.

- Operational efficiency: data facilitates the internal operations of companies by allowing for more streamlined processes that reduce costs and simplify operations. Optimization leads to greater productivity and overall efficiency.

- Innovative product creation: thanks to big data, companies are easily aware of new market trends. This enables them to adapt to the market by creating innovative products and adopting cutting-edge strategies.

- Risk attenuation: data analytics are necessary to reduce the risks a company may incur. Knowing as much information as possible makes it possible to ensure the safety of one's work and to act proactively to ward off risk.

- Improving healthcare and research: big data plays a crucial role in the healthcare industry. With big data analytics, physicians and medical researchers can create predictive models for diseases or ad hoc treatments for each patient, helping scientists achieve important breakthroughs and significant discoveries. Such as the analysis we are going to do in the next chapters about the environmental impact on the population about risk factors.

Big data has become a critical asset for organizations in the modern era. By harnessing the power of big data analytics, businesses can gain a competitive edge, improve decision-making, enhance customer experiences, and drive innovation. It has the potential to transform industries and revolutionize the way organizations operate and succeed in today's data-driven world.

Bruce Johnson, a consultant of Minnesota, once said: "The concept of big data is just that: a concept for the value an organization can realize from in-depth analysis of all data. The concept of big data is, therefore, not a database or data architecture but is more the solutions that leverage any and all data, wherever they come from. In healthcare, the concepts of big data are enabled only in organizations that focus on data capture, management, and usage" [1].

1.3 Risk Social Determinant

One of the most important scopes of the use of big data in health data science for the public research question is to deal with risk social determinants [4]. Social determinants of health are the conditions in which people are born and raised, that affect their health and well-being; they are social and economic factors that increase the likelihood of negative health outcomes. Examples of risk social determinants include poverty, limited access to healthcare, social isolation, exposure to environmental toxins, and unhealthy living conditions. These factors also impact access to healthcare, education, and other resources that can impact health outcomes. These determinants can affect an individual's health outcomes and can vary based on a variety of factors, including race, ethnicity, and socioeconomic status. It is important to keep track of health social determinants in order to improve public health and reduce inequalities.

Enabling risk factors refer to social and material factors that positively impact an individual's ability to improve their health, such as access to education, adequate income, and health insurance, and related lifestyle habits, such as propensity to smoke, alcohol, physical activity, and a healthy, varied diet.

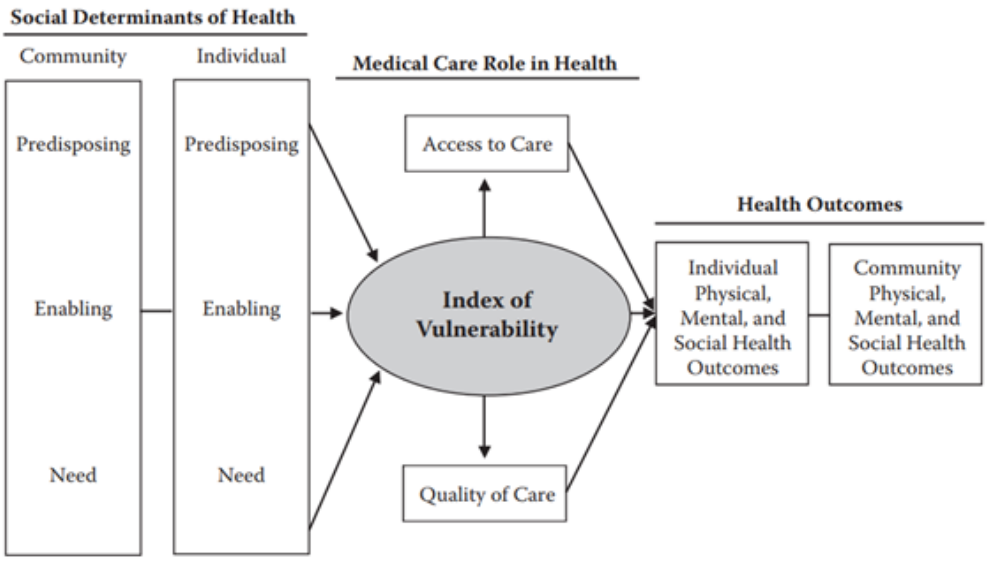

Fig. 1.1: Vulnerability and unmet healthcare needs: the influence of multiple risk factors. Source: Big Data and Health Analytics, by Marconi and Lehmann [1].

Community factors like the availability of well-paying jobs and quality health services also fall under this category. Need risk factors, on the other hand, are existing health issues that can hinder an individual's ability to address other health concerns. These can include personal factors like mental health problems and disabilities, as well as community factors like high rates of communicable diseases, the easy availability of tobacco products, or the presence of liquor stores that encourage alcohol abuse.

Figure 1.1 shows a model that explains how social determinants of health could arise both from the community and individuals. This model organizes social risk factors into three aspects: predisposing, enabling, and need. Predisposing risk factors refer to individual and community factors that predispose individuals to poor health, such as age, gender and ethnicity. Studying risk factors enhances an individual's ability to improve their health, such as income, education, and access to healthcare. Need risk factors are existing health problems that hinder an individual's ability to address other health issues, such as personal mental health issues or disease rates. The model emphasizes the convergence of social risks and calls for an integrated approach to data collection and reporting to identify groups with the highest social risk and prioritize limited resources toward them.

The relationship between health and social factors like poverty and limited access to healthcare implies that interventions that focus on a single risk factor will be less effective than those targeting multiple risks. Public health professionals should develop new data collection strategies to monitor these critical risk factors and implement programs that cross sectors.

For policymakers, this means funding programs that target a range of risk factors common across priority health disparities rather than separate conditions and sub-populations. The general model offered here is a tool for public health, medicine, and policy professionals to understand and address social determinants of health. Big data can also assist in addressing these determinants. From the study and analysis of data surveys, for example, it can be guessed that the set of social risks causes a worsening of the health condition of people. In order to identify the groups with the highest social risk and to direct the limited resources in that direction, it is necessary to focus on the relationship between vulnerability status and hoped-for outcomes.

1.3.1 California Endowment

An example of how analyzing data on social determinants is able to create health policies and adapt services with the aim of improving overall health will be shown below.

A private non-profit company called California Endowment, established in 1996, has promoted an initiative called "Building Healthy Communities" to help individuals with limited access to services in California due to a lower income. This initiative take a period of 10 years, with the aim to provide a healthy home for all children, giving them a healthy diet and the opportunity to attend school, trying to eliminate the violence they have suffered in the past. From its main objectives, it can be easily deduced that the goal is a radical change that will affect the long term. The organization seeks to increase healthy foods, improve air and transportation quality, enhance basic health services, and

ensure greater public safety. To conduct this analysis, the company decided to focus on 14 of California's neediest communities, using big data to choose the neediest and highest-risk settings. The main measures that influenced the choice are poverty, unemployment, school absenteeism rate, and crime rates.

The "Building Healthy Communities" initiative also follows many short- to medium-term goals, such as, for example, developing advocacy activities, and even in these cases, data plays a major role in conducting operations. More than $7,000$ randomly selected individuals from these communities are now included in the biannual Health Profile publications in order to prove how these communities are changing over time thanks to these analyses and suggest ways in which health policies can be improved.

1.3.2 Health Care Surveys and Data Surveillance

To begin addressing social determinants of health in a community, it's important to gain insight into residents' lived experiences. By examining the social factors that impact the community and conducting screenings to identify them, healthcare programs can develop effective strategies for addressing these issues [5]. This initial step can help identify gaps in care and inform targeted interventions to improve health outcomes. For this reason, data collected regarding social determinants are in great part derived from data surveillance and health care surveys. In this section, both topics will be covered and explained.

The presence of health surveys is relevant to improve public health care. The themes covered are different, ranging from patient experience in a hospital or during a disease treatment to the quality of care, personal information on health, demographics, and prevention. All these elements give a view to public health, allowing governments and institutions responsible for health to introduce health policy with the aim of reducing risk factors. The questionnaires are generally based on standardized questions and are not aimed at a specific subgroup of the population, but the goal is to obtain data representing the population sample. The answers are used to identify areas that need improvement in health care and to measure the effectiveness of interventions for the well-being of patients. They also provide details about personal opinions and one's experiences in health care, allowing for the revision of certain health care policies if necessary through the sending and collection of feedback in the area of care. Examples of surveys that have actually interfered in changing statements in the field of health policies are the Survey on Aging, Health and Retirement in Europe (SHARE) and the European Health Interview Survey (EHIS), which through the use of new state-of-the-art tools have helped improve health policies. Finally, the presence of key health policy performance indicators makes it possible to constantly monitor the level of performance and track trends over time.

Speaking of surveys, it is also necessary to introduce the concept of data surveillance. Data surveillance, as well as health surveys, is very important to monitor problems related to public health in order to assess the effectiveness of health policies and regulations and possibly improve them where necessary. In addition, it is essential to identify those areas where the work of the health sector is lacking or completely lacking so that this problem can be solved through targeted actions that aim to improve public health with the help of additional resources or interventions in high-risk geographic areas. For example, it enables healthcare workers to control possible disease outbreaks and the spread of infectious diseases and implement preventative measures to protect public health. Overall, data surveillance is a critical tool for improving the quality of care and ensuring the health and well-being of communities [6]. The surveillance systems can be established in a variety of settings to study various populations and conditions, and effective systems have specific, well-defined objectives, standardized data collection, frequent analysis, and results dissemination to those who need the information.

The data collected include demographic, socioeconomic, clinical characteristics, key outcomes, such as disease complications and mortality, and risk factor data. Data can be collected from population-based surveys or healthcare provider-based systems like sentinel surveillance, which monitors key health events in healthcare settings. Also, patient organizations and medical and health ministry registries collect data for a wide range of purposes. Health ministry registries are national in scope and the most suitable for common diseases and conditions, while medical registries are particularly useful for rare conditions. These data help to monitor disease trends, determine disease patterns across various populations, guide planning and evaluation of disease control programs, allocate health resources, and advance clinical, epidemiological, and health services research.

Data collection is essential for any surveillance system, and it's crucial that the data collection process follows specific principles. The first step is to define clear surveillance objectives that will guide the choice of data elements that have to be accurate and complete. Measurement standards are critical, and case definitions must be clear, with diagnostic data required to validate events for some systems. Standardization of the data collection process is essential for comparing population groups, geographic areas, or trends over time. A good way to do it is by collecting the minimum amount of data required by your analysis and using tools for data collection that are clear and easy to understand. The identification of the proper target population and selection of an adequate sampling strategy is essential if the data are to be representative. Ethical principles, including confidentiality laws, should be followed during data collection, and privacy must be respected. Once data is collected, secure database systems and proper data management and quality control procedures should be in place.

When analyzing surveillance data, cross-sectional models of the population are used to study its outcomes and risk factors, which will, in turn, serve to analyze trends over that period. The goal of the study is to monitor data regarding various diseases to understand the level of prevalence in the population and comprehend whether the number of infected people is increasing or decreasing, as well as the prevalence of the associated risk factors. Once the situation is understood, it is necessary to share all the information gathered with the public, government, and health workers so that they can arrive at optimal resolutions with awareness. The results should be shared widely with the public, healthcare workers, and public health authorities, as well as local and national governments and policymakers, as greater knowledge of risk allows for more informed decision-making. Analyses with a more advanced level are useful to look for possible links that unite risk factors and resulting diseases in order to give rise to interventions aimed at reducing the risk of complications or the spread of a communicable disease.

1.4 Administrative Data and Use of Big Data

The use of big data in private research demand is exploited through analytics in healthcare. Analytics in healthcare is an ensemble of statistical methods used to study the data with the aim of improving the experience of care and the quality of the treatment. For a proper use of analytics, it is necessary for the data collected to be aggregated at the company level to keep informed about patient risk. There are two main brackets of healthcare analytics that can be used to impact in healthcare, described below:

- Utilization analytics: This kind of analysis is called retrospective in that it looks backward at the treatments and diagnoses already made. It specializes strongly in those treatments that have been very costly, and through their study, an attempt is made to devise a strategy that lowers costs but at the same time improves the history. Since it is a past-oriented analysis, one can only deduce which operations not to do instead of deriving a different treatment. However, its usefulness is valuable as it allows not only cost reduction but also inappropriate use of limited resources.
- Predictive analytics: This analysis is prospective. It looks into the future based on new, already-known attempts to predict the probability of future events occurring. For this type of study, nonlinear models are used that not only identify the correlation between cost and risk but also provide insights into the best future use of the data already collected. In fact, given limited monetary resources, predictive analytics is very important to prevent waste and increase the quality of patient care.

It has been noted that there are several areas where, through the use of predictive and utilization analytics, there has been a strong increase in health status and improvement in

savings management. Using analytics in healthcare can lead to significant cost savings, improved compliance, and better health outcomes for patients. For example, analytics can be used in primary care to identify areas of improvement for patients' health and assess their future risk, which can inform plans of treatment in order to motivate patients to conduct a healthier life but, more importantly, to reduce the possibility of future readmission of the patient due to poor care or unhealthy habits: in fact, analytics can be used to identify patients in need of care management programs and health coaching in order to reduce this kind of risk.

1.4.1 The IHI Triple Aim initiative

One application of utilization and predictive analytics in patient care is the "Triple Aim". The Triple Aim was establish by the Institute for Healthcare Improvement (IHI) in response to the high costs and global challenges of healthcare systems, particularly in the US, where healthcare accounts for 17%-20% of GDP.

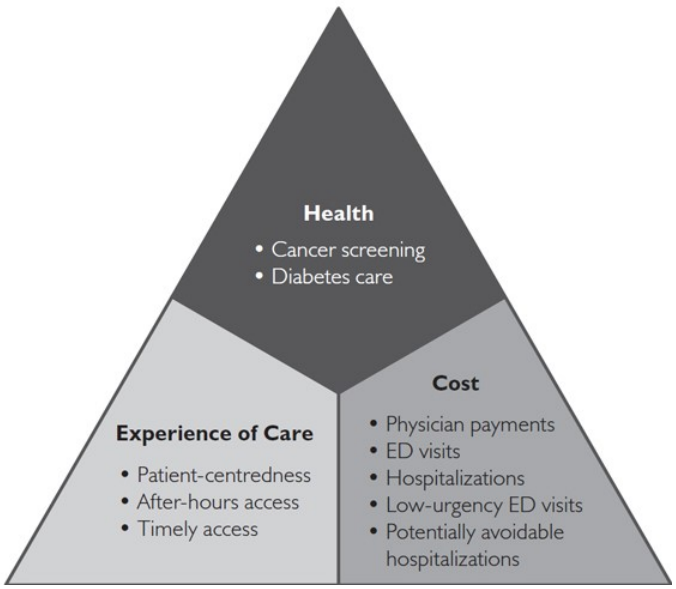

Fig. 1.2: Triangle of the Triple Aim, image taken from [2].

The triple aim is an approach adopted to improve the quality of health system performance based on the achievement of three goals: lower per capita cost of healthcare through improvement, improve individual health, enhance the patient experience of care, and increase quality and satisfaction. These concepts are summarized by Figure 1.2 and are represented in the form of a triangle where the quadrilaterals represent the three main purposes [2].

The IHI advocates for a systematic approach to change by identifying target populations, developing system aims and measures, and testing and scaling up projects

that take into account local needs and conditions. The IHI recommends harnessing a range of community determinants of health, empowering individuals and families, and broadening the role of primary care to provide a seamless journey through a person's life.

Examples of areas where the Triple Aim can be applied include accountable care organizations, bundled payments, new models of primary care, and the introduction of higher-performance technologies. The IHI's innovation team developed a concept design composed of five components, including a focus on individuals and families, the necessary addition of new and efficient primary care services, monitoring and implementation of policies to improve population health, tools that can monitor costs, and system implementation. Organizations that exploit the Triple Aim will have healthier populations, with less burden of illness and much more coordinated care. If the cost per capita of care will be reduce or stabilize, the Triple Aim can give businesses the opportunity to be more competitive and provide more flexibility for public investment in vital community activities.

1.4.2 Machine Learning's Effect on Healthcare Policy: Lessons from Five Case Studies

Artificial Intelligence has made significant contributions to health policy knowledge by leveraging machine learning (ML) [7] to analyze large population-level datasets such as Electronic Health Records (EHRs). This has led to improved predictions for clinical outcomes and a better understanding of disease behaviors. To support this point, five case studies are outlined below.

- Case 1 discusses healthcare fraud and the overuse of tests and medical examinations that are not supported by scientific proof. ML-driven methods can be used to provide intelligent recommendations and validations to patients, and use policies that ensure the validity of clinical recommendations.

- Case 2 focuses on medical tests and areas with limited access to healthcare and suggests that an automated recommendation system could help cover some of these areas. Patients' stratification and classification using clinical and sub-clinical tests could aid in intelligently choosing appropriate tests and lowering health costs for patients.

- Case 3 highlights the lack of preventive practices and evaluates biomarkers concerned with diet to point to attributes that could lead to chronic diseases. Preventive policies are beneficial for the general good of Americans and ought to be implemented.

- Case 4 discusses defensive medicine and how ML methods can provide cheaper solutions to clinical tests. Group-specific recommendations for preventive measures can reduce over-testing and avoid unnecessary expenditures.
- Case 5 presents ML methods that can aid in immunization recommendations and e-access to healthcare. In order to make patients independent in self-assessment and management of their immunization.

1.4.3 Data Protection in the USA

In the USA, there is no federal legislation dealing with data protection but a series of legal instruments at the state level that protect data protection. The protection is, therefore, indirect, as it is based on consumer protection regulations. The most important ones are:

- Federal Trade Commission Act: law passed in 1914 creating the Federal Trade Commission to protect consumers from unfair trade practices;
- Financial Services Modernization Act, to protect consumers in the financial services industry;
- Health Insurance Portability and Accountability Act (HIPAA), which regulates health information used by hospitals, health insurance companies, pharmacies, and their processors;
- Can-Spam Act, which regulates the collection and use of telephone numbers for marketing purposes;
- Stored Communications Act (SCA), recently novated by the Clarifying Lawful Overseas Use of Data Act (CLOUD), which went into effect on March 23 March, 2018, and established that an electronic communications provider is required to share with the federal government even data held on servers located abroad, even if they belong to a foreign entity.

These are complemented by laws in individual states to protect personal data, such as the California Consumer Privacy Act (CCPA), a local legislation intended to improve data and consumer protections for all the residents of the state.

To sum up, in this chapter, a general context has been given on the meaning of health data science, the two main aspects it covers (the public aspect and the individual aspect), and how big data analysis can impact those areas. Despite this, the thesis will focus exclusively not on the individual but on the public, not on data collected by physicians in hospitals but on data collected from questionnaires and data surveillance, a topic that will be explored in more detail in the next chapter.

CHAPTER 2

From Data to Action: Using BRFSS to Understand Risk Factors and Promote Health

Analytics are useful for discovering gaps in prevention and encouraging patients through outreach programs in order to improve the overall health of the population. This chapter will explain the role of big data in healthcare risk and prevention. Big data is very useful in risk and prevention care because allows greater confidence in decision-making through insights from health surveys and data surveillance. In fact, it supports primary prevention strategies by reducing risk factors using modern data analytics tools and technologies and helps insurance providers move from pure risk protection to risk prediction and prevention. Overall, the use of big data in health care can play a crucial role in identifying at-risk populations, preventing disease, and improving health outcomes. For this reason, it will be introduced the Behavioral Risk Factor Surveillance System (BRFSS) survey, a telephone survey that provides estimates on key behaviors related to chronic diseases, injuries, and preventive health practices.

2.1 What is Disease Prevention?

The early detection and prevention of disease play a crucial role in reducing healthcare costs and deaths. According to the article “Big data analytics for preventive medicine”

[8], there is a significant number of preventable deaths in the world, and the need for affordable interventions to reduce healthcare costs. The medical costs expense over a person's lifetime has risen significantly in the past years, starting from $14,000$ dollars and going up to $83,000$ dollars per individual. The percentage of people who suffer from mortal diseases, including diabetes, cardiovascular diseases, and cancer, is rising, being the majority cause of death worldwide. The primary goal in the healthcare sector is to reduce the number of risk factors and costs while reducing the prevalence of diseases and increasing the importance of prevention. In fact, it is essential to develop cost-saving and cost-effective preventive solutions to reduce economic burdens. Traditional preventive measures could lead to many healthcare benefits, but they are also balanced by the intervention costs.

Prevention does not only consist of expedients to be taken by humans in regard to the population, such as reducing smog and pollution, but also of actions at the individual level, such as following of diet, taking constant exercise, taking care of one's hygiene, and so on. Disease prevention is a critical aspect of public health and can help reduce the burden of disease on individuals and society as a whole. Disease prevention actions could be summarized by a three stages methodology:

- Primary prevention involves reducing the occurrence of new cases through risk factor reduction, health promotion, and other protective measures, such as encouraging healthier lifestyles.
- Secondary prevention aims to cure or slow down the disease's progress by detecting it early and reducing its impact, such as through lifestyle modification or screening for pre-clinical pathological changes.
- Tertiary prevention aims to enhance the patient's quality of life through therapy and rehabilitation once the disease has been treated in its acute clinical phase.

Effective implementation of all three stages helps improve the quality of life and can reduce unnecessary healthcare costs. Medical knowledge is required to predict the probability of individuals contracting a disease.

So, how can big data help to reduce diseases and mortality in the population? One of the most important roles of health data science is the collection and study of public data to increase prevention and combat risk factors. There are many research institutions that focus on prevention by leveraging data to identify risk areas and intervene in the most efficient way to improve public health and prevent hard-to-treat diseases. Data science approaches show promise in addressing health needs, such as identifying high-risk patients and improving workflow optimization, disease surveillance, and cost

control. Overall, health data science provides a powerful tool for preventing diseases and improving population health by identifying risk factors, developing interventions, and optimizing healthcare processes. Working on preventable risk factors, both at the individual and community levels, is the first step in contributing to prevention using big data. Individuals monitoring themselves (quantified self) and taking action to improve their own health enables data collection that aggregates to the population and institutional level, thus improving public health.

In the article "Big data and disease prevention: From Quantified Self to Quantified Communities" [9], two ways are explained to manage the issue: the first one is to prevent diseases by increasing our understanding of the social, behavioral, and environmental factors that affect health at the population level, the second way is to enable better targeting of disease prevention efforts towards specific sub-populations through precision prevention.

Prevention first requires identifying risk factors such as smoking, alcohol, and diet and then planning how to intervene to improve health. However, public health still holds too little data, too little geographic information, and time jumps that hinder analysis.

Most of the data is collected through health apps such as apps that offer exercise, diets, and other checks. All this data is collected and studied in order to bring improvements in the field. How does big data accelerate data processing? First, big data provides massive datasets that allow for analyses at the population level, sub-population level, and personal level. This means that new risk factors can be discovered, and personalized risk factors can be identified for individuals, which will help people to prevent diseases more effectively. For example, some dietary nutrients may be beneficial for some people but harmful to others, and only with large datasets can we detect these differences and provide personalized nutrition advice. Second, new passive sensors can be used to collect richer, more detailed data on potential risk factors for longer follow-up periods than standard epidemiological questionnaires. This will help extract new insights from big data and improve our understanding of potential risk factors for diseases. Some examples of this online collection of data are PatientsLikeMe and the Health Tracking Network, which collect data on chronic diseases, and Google Flu Trends to track and predict influenza outbreaks [9].

By monitoring an individual's health behavior and providing real-time feedback on their performance in relation to personalized targets, it is possible to support them in achieving their behavior goals. Ideally, behavioral data should be passively collected to allow continuous, long-term follow-up without requiring additional effort from the patient. Algorithms can provide personalized feedback to assist with behavior modification at key moments of decision-making, setting this application of big data apart

from traditional public health uses of behavioral or health data.

The article "Beyond Health Care: The Role of Social Determinants in Promoting Health and Health Equity" [10] explains that in addition to providing coverage expansions, the Affordable Care Act's (ACA) Medicaid expansion has opened up new avenues for initiatives aimed at addressing social determinants of health, particularly the fact that social factors are becoming increasingly important in health care decisions and analyses. While a broad range of initiatives have been launched within the healthcare system and within Medicaid specifically, many challenges remain. Funding streams and service capacity may not align, and there are gaps and inconsistencies in data on social determinants of health. Despite centers for Medicare and Medicaid Services' assertion that work requirements in Medicaid will help individuals attain independence and improve health outcomes, eligible individuals may face challenges fulfilling these requirements and lose Medicaid coverage as a result.

2.1.1 Global Burden of Diseases

In 1992, the World Bank requested a study called Global Burden of Disease (GDB) [11] to be conducted, which was undertaken with the full participation and collaboration of the World Health Organization (WHO) over the course of 4.5 years. The study identified leading risk factors contributing to deaths and disability, including dietary risks, high blood pressure, and smoking. On the Institute for Health Metrics and Evaluation (IHME) website, academic journals and publications contain the results of all GBD studies.

The GBD study aimed to achieve four specific goals:

1) To generate estimates of mortality, broken down by cause, age, and sex, for the world that is divided into eight geographic regions (North America, Western Europe, Asia Pacific, Australasia, Latin America, and Caribbean, Sub-Saharan Africa, South Asia, and North Africa and the Middle East) and based on 107 major causes.

2) To generate estimates of the incidence, prevalence, duration, and case-fatality for 483 disabling sequel resulting from the above causes, broken down by age, sex, and region.

3) To estimate the fraction of mortality and disability attributable to 10 major risk factors, broken down by age, sex, and region.

4) To create and evolve cases and study of disability and mortality, originating by country, sex, and age.

Diseases and injuries in the GBD were divided into three clusters: Group 1, consisting of communicable diseases, side effects of pregnancy, conditions derived in the perinatal period, and nutritional deficiencies; Group 2, encompassing the non-communicable diseases; and Group 3, comprising all injuries, whether intentional or unintentional.

The GBD informs decision-making and improves population health by identifying leading risk factors contributing to deaths and disability and helping track progress toward health-related Sustainable Development Goals (SDGs), an ensemble of 17 goals to achieve before 2030 to end poverty and protect the planet. The methodology for calculating the "Burden of Disease" (BOD) is an extremely complex methodology with many facets [12]. The key part of the methodology is to create as detailed a database as possible on the descriptive epidemiology of the most important disease conditions that cause death and disability in the area of interest. Starting from the data in such a database and using the data available in the consistency of the survey and the opinions of qualified experts in the individual diseases, through appropriate statistical techniques, certain disease-related characteristics, such as incidence and prevalence, are estimated. A prerequisite of the whole methodology is the possibility of assessing health states in the sense of assigning them numerical ratings or "weights" (or alternatively using ratings of weights already made by others), which measure the relative degree of quality of life in the disease condition, called disability weights.

The GBD uses as a measure of the impact of a particular disease a unit called DALYs (Disability Adjusted Life Years), which is the sum of years of life lost due to premature mortality (Years of Life Lost, YLLs) and years of life lived in sub-optimal health or disability (Years of Life lived with Disability, YLDs). To calculate the years lived with disability (YLD), scientists use the disability weights, taking values from 0 to 1 that indicate the health status, where 0 stands for perfect health and 1 for bad health. Disability weights can vary depending on the specific health condition, the population being studied, and the methodology used to determine them. The formula to calculate this measure for a specific disease is the following:

$$\mathrm{YLD} = \mathrm{I} \times \mathrm{DW} \times \mathrm{L}$$

where I is the number of incident cases of that disease, DW is the disability weight of that disease, and L is the average period of time from the case onset to either remission or death (in years). Instead, the Years of Life Lost (YLL) for a specific disease is computed as:

$$\mathrm{YLL} = \mathrm{N} \times \mathrm{E}$$

where N is the number of deaths due to that specific disease and E is the standard life expectancy at the age of death. Finally, the DALY metric is simply computed as YLD + YLL [13].

Fortunately, the GBD study is a critical resource for policymakers and public health professionals to effectively prioritize health interventions and allocate resources. This study provides a wealth of data on the world's current health status, which can help identify the most urgent health concerns impacting different populations globally.

2.2 Health Care Policy in the USA

Since the source of the dataset that will be analyzed in the next chapter is American and all the data are collected from American citizens, it is onerous to briefly introduce the healthcare system and data collection policy in this country.

In America, the health care medical system is financed both by government agencies that draw on public sources, the most important of which are Medicare and Medicaid, and by private insurance: the system is a mix of public and private insurance, this causes partial coverage that cannot fully cover health care needs. In fact, healthcare spending is the highest in the developed world, but there are low rates of preventive healthcare utilization: for example, only 8% of Americans obtain routine preventive screenings. Moreover, less than half of USA states have average or above-average metrics for preventive healthcare. Most healthcare research focuses on disease treatment rather than prevention, but prevention should be a national priority, especially since statistical methods could be used to improve preventive practices and policies.

In the USA, there are a huge number of initiatives with the aim to promote the use of data analytics and other technologies in healthcare for the collection and analysis of data, such as the Affordable Care Act and the Precision Medicine Initiative [14]. The Affordable Care Act is a healthcare reform law signed in 2010 with the goal of expanding access to affordable health insurance coverage to more Americans, and the Precision Medicine Initiative, launched by former President Obama in 2015, is a research effort designed to revolutionize medicine by studying the genomic, lifestyle, behavioral, and environmental factors that contribute to individual health and diseases. The initiative seeks to develop more precise treatments that take into account each person's unique characteristics. Also, several organizations are working to promote the adoption of big data technologies in healthcare with the aim of improving the quality of care and reducing costs through the use of big data.

In the United States, the good health index is very low despite per capita expenditures on health care being the highest among all nations. In fact, nearly half of all deaths are

caused by risky behavioral factors such as tobacco, physical and social environment, and diet. The major cause of premature death is preventable chronic diseases, such as stroke, heart disease, and cancer. Currently, 10% of Americans rate their health as fair or poor, and 36% of adults are considered obese. Over the next 20 years, Americans are projected to suffer from about 8.5 million new cases of diabetes, 7.3 million cases of heart disease and stroke, and over $660,000$ cases of cancer, potentially costing up to dollar 66 billion per year. One of the factors that caused this type of situation is the difficulties in accessing healthcare providers. [15].The possibility of access, like an economic attribute, has a high correlation with poverty rates, education levels, and job rates. Socioeconomic factors play a significant role in the inadequate management of diseases and hinder patients from seeking early screening and preventive care. In terms of healthcare policy, access to medical care is limited by multiple factors, including also the low number of doctors in individual regions of the country. In the US, there were 2.6 medical doctors per 1000 patients in 2013; the doctor-to-patient ratio is below the average estimated by OECD-recommended that is 3.3 physicians per 1000 patients. Furthermore, market research projections from the OECD suggest that an additional $52,000$ primary care physicians will be needed by 2025 to meet demand, especially given the ACA's provisions for increasing the number of physicians. While the U.S. average for active patient-care doctors per $100,000$ people stood at 225.6 in 2014, there are significant differences between states. Massachusetts has the highest rate of active doctors, with 349.5 active doctors per $100,000$ people (the ACA is a revival of the Massachusetts healthcare system), while Mississippi has only 170.3 doctors per $100,000$ people.

To sum up, in America, there are many policies to protect data and organizations and initiatives that collect them in order to improve collectible health, but at the same time, collecting data from public health surveys is really important to treat public health. USA, compared to other states, is the most likely open to publish the data collected from the questionnaires, allowing users to have greater insight into the health status of the population. Thanks to this choice, it is possible to have full access to BRFSS, the data system that will be analyzed in the next chapter.

2.3 BRFSS: America's Leading Health Survey System

The Behavioral Risk Factor Surveillance System (BRFSS) [16] is the main national health-related telephone survey that gathers data about USA risk behaviors, the major conditions of chronic disease, and the exploitation of the forecast in order to improve public health. Starting with 15 member states in 1984, nowadays, BRFSS collects data from all 50 U.S. states and other few territories such as Columbia, with more than $400,000$ adult interviews conducted annually. the time duration of the survey

is 25 minutes. It may be less or more time, depending on the number of questions asked and how long you take to answer. As the largest continuously conducted health survey system in the world, the collection of behavioral health data at the state and local levels has made BRFSS a potent tool for directing and advancing health promotion activities. Consequently, users of BRFSS have demanded additional data and survey questions. Presently, BRFSS has broad sponsorship, including most divisions in the Centers for Disease Control and Prevention (CDC), the National Center for Chronic Disease Prevention and the Health Promotion, as well as federal agencies, such as the Mental Health Services Administration, the Department of Veterans Affairs, and the Health Resources and Services Administration. Moreover, BRFSS staff has received technical assistance requests from countries interested in establishing similar surveillance systems. The background and evolution of BRFSS can be traced back to the 1970s when the CDC developed what was known as the BRFSS Pilot Project. This pilot project was an effort to determine the feasibility of conducting behavioral surveillance and involved initial surveys conducted in 29 states.

The BRFSS system has evolved over the years to become a critical tool for understanding and monitoring health trends in the USA and beyond. Some key milestones and developments in the history of BRFSS include:

- The introduction of cellular phone sampling in 2011 to address declining response rates for landline surveys.
- The addition of state-added questions to the core questionnaire in 1993 allowed states to collect more specific data on their populations.
- The inclusion of modules on emerging public health issues such as HIV/AIDS, asthma, and diabetes in the 1990s and 2000s.
- The implementation of a new sampling strategy in 2017 known as the "combined landline and cell phone" approach.

The methodology of BRFSS involves sampling strategies and questionnaires designed to collect data on health risk behaviors. To collect data for BRFSS, telephone interviews are conducted using a stratified sampling design with random-digit-dialing that ensures the randomness and fairness of the surveys. The BRFSS system collects state-specific data on health risk behaviors using a stratification method to weight the data. The types of data collected by BRFSS include information on various health risk behaviors such as tobacco use, physical activity, diet and nutrition, chronic health conditions, and access to healthcare. Therefore, the BRFSS methodology and design includes data collection through telephone interviews using random numerical selection and disproportionate

stratified sample [17]. Random numerical selection or Random-digit-dialing (RDD) is a method used in health surveys that, through probabilistic sampling, selects the participants by generating a random phone number. Moreover, a disproportionate stratified sample design (DSS) is a stratified method that divides the population into two groups, or strata. The first group is the high-density stratum, refers to that number that in the rubric more than one listed household numbers in its hundred block. A hundred block refers to a group of 100 telephone numbers that share the same area code, prefix, and the first two digits of the suffix, along with all possible combinations of the last two digits. This group is sampled at a higher rate respect the second one, to have more household respondents. Instead of, the second groups, with low density stratum, is supposed to refers to fewer households [14].

The sampling strategies and questionnaire design are intended to collect uniform state-specific data on health risk behaviors; for this reason, the type of data collected includes information on various health risk behaviors and access to healthcare. It is important to remember the limitations of this data: being data self-reported, errors can be frequent as most participants do not work in the health or medical field, and also could be some bias influenced subjectively by their own health status. Thus, in analyzing and studying these data, it is important to take into account the nonexistence of absolute truthfulness in the data collected due to its nature.

2.3.1 Assessing the Reliability of BRFSS

The article "A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS)" [18] analyzed the literature and identified 32 examples of reliability and validity tests of the BRFSS published between 2004-2011 across ten categories. However, the literature did not evenly cover all topics in the BRFSS, and quality varied among publications. In fact, the largest number of articles were associated with physical activity, access to healthcare, immunization, and preventive testing for chronic diseases/conditions. Through this research, it was concluded that:

- Self-reported diagnoses of chronic conditions through BRFSS and other surveys showed high reliability. Most test/retest research showed high levels of reliability, although some questions' reliability worsened with longer periods between questionnaire administrations.

- The validity of self-reported data was best for discrete events such as immunization or testing and behaviors that were not sensitive. Rankings produced through the

categorical question correlated with access to healthcare, immunization, preventive testing, screening, and question on disease or chronic condition.

Differences between the BRFSS and other surveys were mainly due to differences between self-reports and physical measures [18]. The BRFSS has been found to be comparable to other national and state-level surveys that investigate similar topics, with few data quality differences noted when compared to a review of literature conducted over a decade ago. The responses are different among the respondents due to several motivation, including differences in survey mode, the number of questions, question wording, and sampling. Even when prevalence rates differed between surveys, other statistical relationships within the datasets remained the same. The veracity of self-reports of chronic conditions compared with physical measures may be due to respondents' lack of awareness of their current presence. The review suggests that continued research on reliability and validity in health fields is necessary, especially with the release of mobile phone data and new weighting methods.

2.3.2 Safeguarding Data Privacy in the BRFSS

The BRFSS policy on data protection and privacy is very strict. The data collection, in fact, takes data protection very seriously. To ensure that the data study is ethical and respects the privacy of participants, some of the 50 USA states require researchers to receive a copy of the institutional review board's approval. The Institutional Review Board (IRB) [19] is a governing body that safeguards the rights and well-being of human research participants, overseeing and evaluating research projects involving human subjects to ensure they adhere to scientifically sound practices, ethical standards, and applicable regulations. The IRB has an important impact even in health surveys, reviewing the protocol of the survey to assess the reliability of data protection. Furthermore, some data are not provided to the public, such as zipper code level data, and also uses national weighting techniques, such as raking, an iterative method used to adjust the weights of population features, such as age, gender, and education. This helps to ensure that the survey results are representative of the target population and that private responses cannot be linked to specific individuals.

BRFSS follows industry-standard practices that consist of collecting data using particular measures with the aim of protecting the confidentially of the survey, assessing authorized personnel, and using suppression norms to ensure the safeguarding of individual-level information. Other recommended approaches might involve setting up committees to assist in determining data sharing with external entities, establishing infrastructure capable of handling the growing amounts of collected data, and implementing robust authentication and access management systems to oversee user

identities and permissions. Adhering to the General Data Protection Regulation (GDPR), an ensemble of data protection and privacy laws created by the European Union, also necessitates that organizations adapt their data collection, storage, sharing, and deletion practices, with non-compliance carrying the potential for significant penalties. These practices collectively aim to ensure the privacy of survey respondents and secure the confidentiality of their data [20].

CHAPTER 3

Behavioral Risk Factor Surveillance System Data Analysis

After discussing the great impact big data has in health care and how data surveillance and health survey results impact health policies, this chapter will analyze the Behavioral Risk Factor Surveillance data focusing on the 2019 year. The analysis starts with an exploration of the data and variables in the dataset with a variety of visualizations to explore the distribution, relationships, and patterns within the data, followed by pre-processing steps undertaken to clean and prepare the data for analysis and then investigate the relationship between risk factors, socio-demographic variables and cancer considering the intrinsic spatial geo-location of the data, through a spatial cluster statistical analysis.

The R programming language is used. This programming language is used particularly for statistical calculations and implementation of analysis techniques in the health care field [21].

3.1 Research Question

The research question focuses on the study of the relationship between selected risk factors, socio-demographic variables, and cancer in the American population. For that,

the source of the data being analyzed is the BRFSS, a health survey that collects annual demographic data, data pertaining to risk factors, data on the spread of disease, and data pertaining to prevention. The state of origin is also specified for each citizen, which is very useful information for the spatial clustering analysis that will be used to identify spatial patterns composed of combinations of risk factors, socio-demographic variables, and cancer. The spatial cluster analysis that will be used is the one proposed by Assunção and coauthors called `skater` (Spatial C(K)luster Analysis by Tree Edge Removal) based on graph theory, in particular, it used the concept of minimum spanning tree cluster analysis [22]. We aim to show how certain social, demographic, and health situations characterized particular areas of the USA.

3.1.1 Data

The datasets used are four:

(a) "BRFSS.ASCII": The dataset following data cleaning and extrapolation of the variables of interest consists of 300,458 observations and 20 variables covering demographic data (such as age, sex, income, education level, and diet), data pertaining to physical and psychological health, data on health insurance, and data on risk factors. The time period considered is 2019; it was decided to study the data prior to 2020 since interference caused by the arrival of Coronavirus on the actual data values is not desired. There are 48 USA states considered: observations derived from Hawaiian Islanders and Eskimo Islanders will not be considered within the analysis as they were removed so that they would not negatively interfere with the spatial analysis being spatial distant to the other state. The source of this dataset is the web page of the Center of Disease Control and Prevention about BRFSS "Survey Data and Documentation: `https://www.cdc.gov/brfss/annual_data/annual_2019.html`.

(b) "Fips.csv": in the "BRFSS.ASCII" dataset, the states are codified in number, and there are not the states's names information. For this reason, the file "Fips.csv" is merged with the "BRFSS.ASCII" dataset in order to create a dataset with a variable state that contains the name of the state and not only the numerical code. It contains 53 observations, one for each of the 53 countries related to the US, including Porto Rico and Guam (that are not present in the "BRFSS.ASCII" dataset), and 2 variables: one contains the numerical code (called fips) and the other the name of the US state. The source of this data is the website GitHub: `https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt`.

(c) "States.shapefile": It is a `shapefile` which contains the location, line, geometry,

and polygon features of the USA states. It is a geospatial vector that is essential to represent the socio-demographic variables (for example) analyzed in a geographical map and to compute spatial clustering. The source of this data is the website ArcGIS Hu: `https://hub.arcgis.com/datasets/1b02c87f62d24508970dc1a6df80c98e/explore`.

(d) "2019-column-layout.csv": Before starting to work on the dataset, it is necessary to use this CSV file containing information about the combined landline and telephone multiple data variable layout of variables that are going to be analyzed in the "BRFSS.ASCII" dataset. The file has three columns: the first one, the `Starting Column`, denotes the index of the column in the "BRFSS.ASCII" file in which the variable is reported; the second column, the `variable name`, contains the name of that variable in BRFSS file and the last column `Field Length` described in how many columns the variable is contained in ASCII file. The dataset is used in the creation of a code that calculates the actual widths of the fields in the ASCII file, widths that are used to read the file into `R`. In the ASCII file, there are multiple subsequent columns that have no independent width of their own and also blank columns between some fields. To have a dataset that can be imported into `R`, it is necessary to remove the group fields and to calculate the effective field widths. The source of this data is the website of the Centers For Disease Control and Prevention: `https://www.cdc.gov/brfss/annual_data/2019/llcp_varlayout_19_onecolumn.html`.

3.2 Exploring the Depths: BRFSS Data through Explanatory Data Analysis

In this section, there will be represented the explanatory data analysis used to analyze and interpret the relationships, patterns, and trends within the risk factors, socio-demographic variables, and cancer. The raw data is derived from the "BRFSS.ASCII" file that contains all the anonymous answers given by American citizens; once the file is downloaded and the data unzipped, the dataset created from the ASCII in `R` is composed of 407607 observations and 280 columns.

From the 280 columns, 21 are selected in this project, and they could be divided into three main groups:

- Demographic variable: provide information about the composition of the population such as the state of provenience and sex (`state`, `SEX` columns), the age variable is divided into six levels (`AGE` column), variable about marriage status (`MARITAL` column), the type of work of the individual (`EMPLOY` column), the level

of education attained by an individual (EDUCA column), and the ownership or not of health coverage (HEALTH.COVERAGE columns).

- Risk factor variable: these variables signal the presence of individuals' risk factors such as smoking, alcohol, weight, body mass index, and exercise (SMOKDAY, AVEDRNK, BMI, WEIGHT, EXERCISE columns, respectively).

- Health variable: a variable that describes the spread of the cancer among the states (CANCER columns), the physical and mental health of an individual (PHYSHLTH, MENTHLTH columns). These last two variables are only mentioned in the thesis and will not be further analyzed.

Before starting the analyses, all missing values were processed. Once the observations equal to NAN were identified, which were very high for the variables WEIGHT, EXERCISE, and BMI, they were been managed and deleted. For most of the variables analyzed, there were frequent refused answers that were indicated in the questionnaire as the numbers 9/ 7, 77/99, 777/999, or 7777/9999 according to the type of variable. These data were also eliminated from the dataset. The numerical variable AVEDRNK has been transformed into two-level categorical variable ALC.class, where 1 refers to not drunker and 2 refers to drunker. Also the numerical variables PHYSHLTH, MENTHLTH have been transformed into three-level categorical variables PHYSHLTH.class, MENTHLTH.class. The question that refers to the two health variables is "Now thinking about your physical/mental health, for how many days during the past 30 days was your physical/mental health not good?", where from 0 to 3 day the level is " Low day n.", from 4 to 9 the level is "High day n." and more than 9 days is "Very high n."

Table 3.1 shows the output of the skimr R function about the factor variables. The information represented is the variable name, the unique value of each variable, and the top counts. The last information reports the two levels of the categorical variable with the highest absolute frequency. From an initial analysis, it is possible to observe among the respondents that: the number of those who have health insurance is higher than those who do not, most of them are college graduates, the number of older adults is higher than the number of young adults, and people with weight problems are many more than those who do not. As for health, the number of people who have been sick several days either physically PHYSHLTH.class or mentally MENTHLTH.class in the past month is lower than those who have not.

Instead, Table 3.2 represents the contingency table of males and females who participated in the survey for each state. It is possible to observe that the states with a higher absolute frequency of respondents are Nebraska (12001) and Maryland (11981).

Table 3.1: List of dataset variables, unique level for each variable, and the top counts of the two values with the higher frequency.

Variable	N. unique	Top counts
HEALTH.COVERAGE	2	**Yes**: 258160, **No**: 21094
CANCER	2	**Yes**: 28630, **No**: 250106
EDUCA	6	**College graduate**: 114035, **Attend college**: 79348
MARITAL	6	**Married**: 148597, **Never Married**: 44570
EMPLOY	8	**Employed**: 121495, **Retired**: 83935
INCOME	8	**More than 75k** : 100616, **From 65k to 75k**: 45787
SMOKEDAY	2	**No**: 239017, **Yes**: 39719
EXERCISE	2	**Yes**: 207299, **No**: 71437
SEX	2	**Female**: 145550, **Male**: 133186
AGE	6	**More than 65 y**: 98861, **From 55y to 64y**: 58842
BMI	3	**Obese and overweight**: 193627, **Norm-weight**: 81015
MENTHLTH.class	3	**Low day n.**: 230513, **Very high day n.**: 34421
PHYSHLTH.class	3	**Low day n.**: 226748, **Very high day n.**: 39215
ALC.class	2	**No**: 240620, **Yes**: 38116

Table 3.2: Absolute frequency distribution for state and sex.

Teste	State	Male count	Female count	Total
1	ALABAMA	2019	2759	4778
2	ALASKA	1081	1411	2492
3	ARIZONA	2856	3160	6016
4	ARKANSAS	1515	1974	3489
5	CALIFORNIA	4357	4042	8399
6	COLORADO	3239	3236	6475
7	CONNECTICUT	2918	3132	6050
8	DELAWARE	1126	1307	2433
9	DIST OF COLUMBIA	807	959	1766
10	FLORIDA	4872	5484	10356
11	GEORGIA	2010	2366	4376
12	GUAM	861	922	1783
13	HAWAII	1295	1104	2399
14	IDAHO	1811	1837	3648
15	ILLINOIS	2102	2308	4410
16	INDIANA	2718	2884	5602
17	IOWA	3457	3592	7049
18	KANSAS	3989	3992	7981
19	KENTUCKY	2161	2431	4592

Continue on the next page

Table 3.2: Absolute frequency distribution for state and sex.

Teste	State	Male count	Female count	Total
20	LOUISIANA	1470	1706	3176
21	MAINE	3410	4124	7534
22	MARYLAND	5560	6411	11971
23	MASSACHUSETTS	2356	2498	4854
24	MICHIGAN	3588	4054	7642
25	MINNESOTA	5558	5303	10861
26	MISSISSIPPI	1475	2060	3535
27	MISSOURI	2403	2689	50922
28	MONTANA	2498	2366	4864
29	NEBRASKA	5736	6265	12001
30	NEVADA	991	1002	1993
31	NEW HAMPSHIRE	1862	2041	3903
32	NEW MEXICO	2150	2283	4433
33	NEW YORK	4279	4416	8695
34	NORTH CAROLINA	1399	1462	2861
35	NORTH DAKOTA	2168	2102	4270
36	OHIO	4029	5037	9066
37	OKLAHOMA	1893	2321	4214
38	OREGON	1980	2027	4007
39	PENNSYLVANIA	2510	2439	4949
40	PORTO RICO	1734	2742	4476
41	RHODE ISLAND	1811	2059	3870
42	SOUTH CAROLINA	2275	2535	4810
43	SOUTH DAKOTA	2143	2373	4516
44	TENNESSEE	1914	2179	4093
45	TEXAS	3689	4237	7926
46	UTAH	4495	4100	8595
47	VERMONT	2089	2330	4419
48	VIRGINIA	3215	3588	6803
49	WASHINGTON	4554	4526	9080
50	WEST VIRGINIA	1828	2145	3973
51	WISCONSIN	1761	1928	3689
52	WYOMING	1545	1817	3362

3.2.1 Demographic Variables

This subsection will show descriptive statistics about the demographic variables, including measures to provide a comprehensive summary of the data for the purpose of getting a general overview of the analyzed dataset.

The AGE variable is a factor variable with six levels, one for each group of age, that describe the distribution of age in American citizen respondents. In Figure 3.1, the x-axis represents the age levels, while the y-axis represents the absolute frequency or percentage of individuals falling within each bracket. The first bracket is about respondents with imputed ages between 18 and24 years of age, the second is between 25 and 34 years of age, the third is between 35 and 44 years of age, the fourth is between 45 and 54 years of age, the fifth is between 55 and 64 years of age and the last one is for respondents with imputed ages between 65 and 99 years of age. It is possible to notice that more than 50% of the observations, the respondents are older than 55 years, and the percentage of very young adults is quite low, about only 5%.

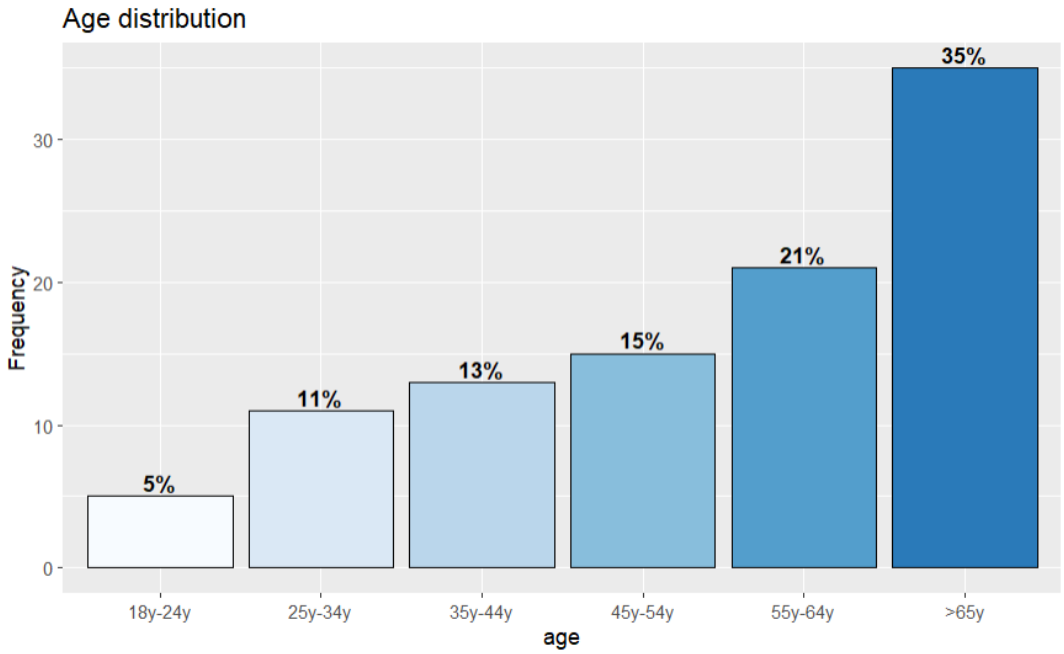

Fig. 3.1: Age distribution of the respondents divided into six levels, in absolute frequency.

From the Figure 3.13 is possible to observe the distribution of old adults relative frequency among the states, where a state with a dark color refers to a state with a smaller number of respondents who are older than 45 years, such as Utah (around $59,02\%$) California (around $59,80\%$). Instead, most states are represented by lighter colors, like Main (around $81,55\%$) and Minnesota (around 80.89%), where there is a higher number of respondents who are older than 45 years.

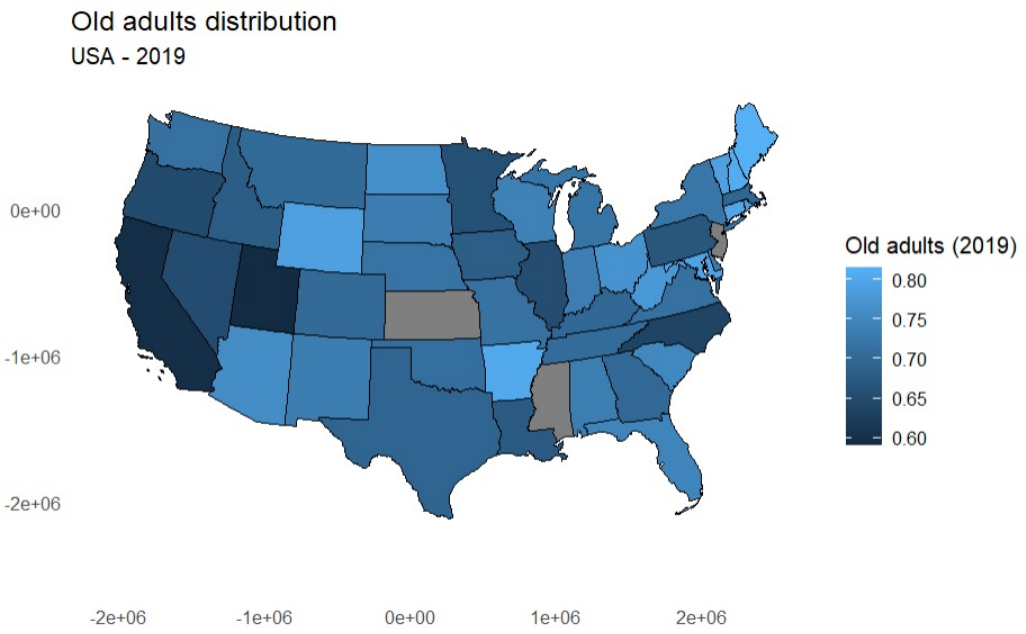

Fig. 3.2: Old adult distribution for each state in relative frequency.

Figure 3.3 illustrates the distribution of the INCOME variable within the BRFSS sample, showcasing the varying levels of earnings divided by age and answer to the question "Is your annual household income from all sources—". The x-axis represents the six different ranges of age, while the y-axis represents the percentage of income for each age bracket. The income brackets are the following: level 2 refers to respondents that earn in a year from $10,000$ to $15,000$ dollar, level 2 earn from $15,000$ to $20,000$ dollar, level 3 from $20,000$ to $25,000$ dollars, level 4 from $25,000$ to $35,000$ dollars, level 5 from $35,000$ to $50,000$ dollars, level 6 from $50,000$ to $65,000$ dollars, level 7 from $65,000$ to $75,000$ dollars and the last level refers to respondents who earn more than 4 dollars per year. From the graph, it is possible to say that individuals ranging in age from 34 to 54 have a higher frequency of the highest income range, while those 18-24 years old have the highest percentage of the lowest income range. Overall, individuals earning more than 75 thousand euros per month have the highest frequency compared to all others.

Analyzing the distribution of the education variable among the BRFSS sample, Figure 3.4 show the six different types of education levels and the absolute frequency for each group of age divided by SEX. The levels of the EDUCA variable that answer the question "What is the highest grade or year of school you completed?" are the following: 1 stands for never attended school, 2 stands for elementary, 3 stands for attending some high school, 4 stands for high school graduate, 5 stands for attending some college or technical school, and 6 stands for college graduate. It is possible to notice that for each age class, the majority of respondents have high school or college degrees. From the sex

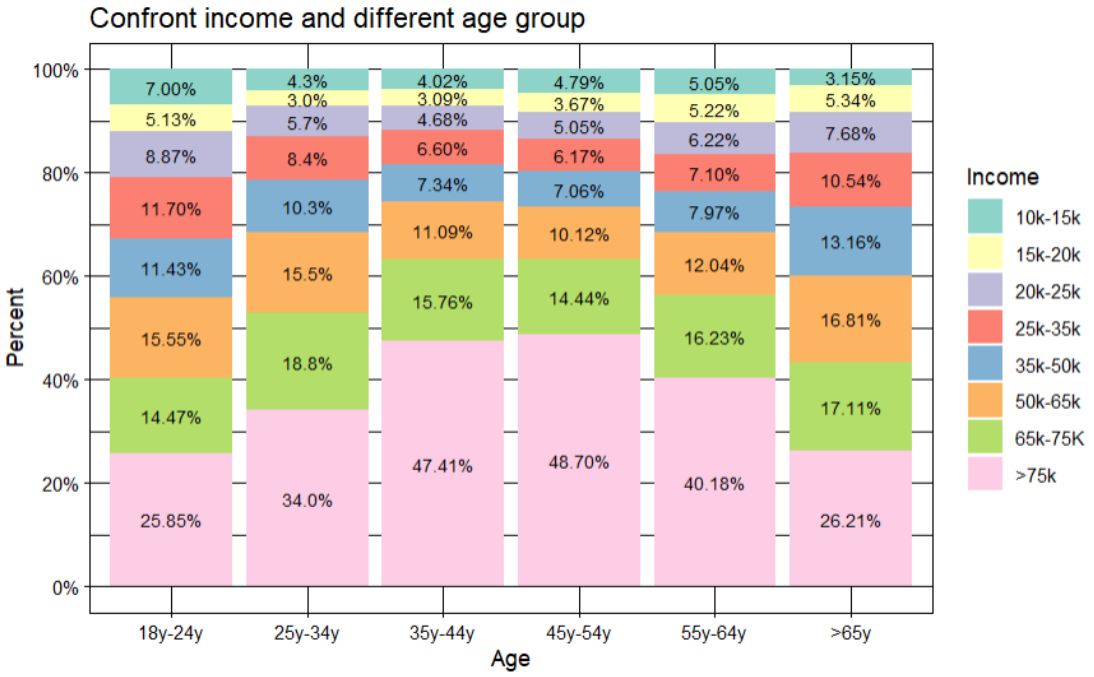

Fig. 3.3: Income distribution for age group, conditional relative frequencies per age

point of view, except for the last age group, the frequency of females who have college graduates is higher than the frequency of males one.

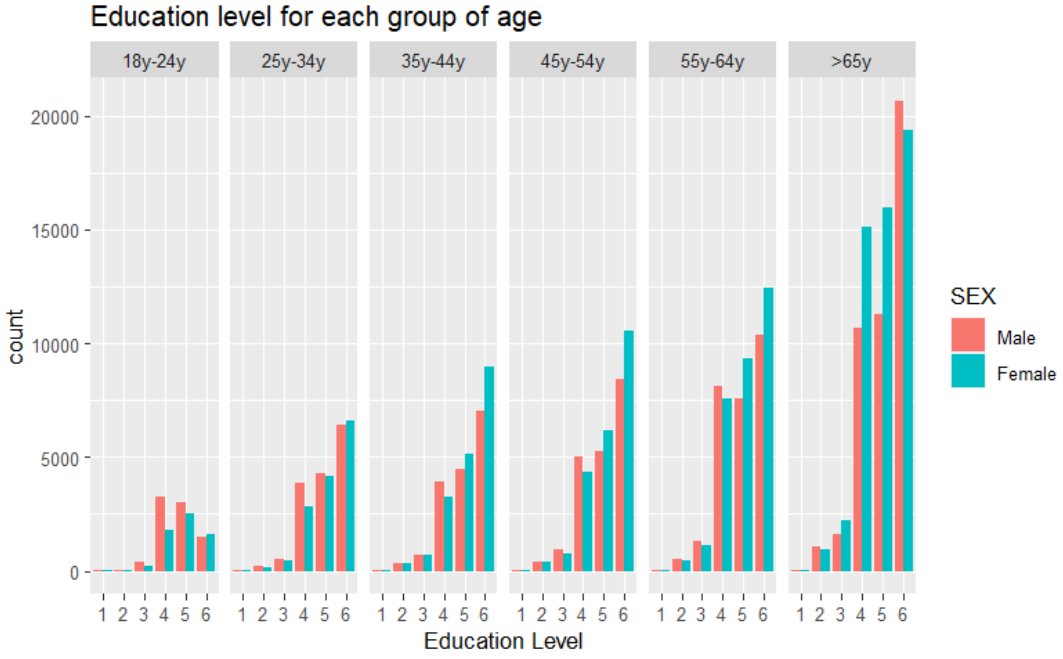

Fig. 3.4: Education level distribution for each age, in absolute frequency, where 1 is Never attended school, 2 is elementary, 3 is attending high school, 4 is High school graduate, 5 is college or technical school, and 6 is for college graduate.

The EMPLOY variable has eight categories to describe the type of job of the BRFSS

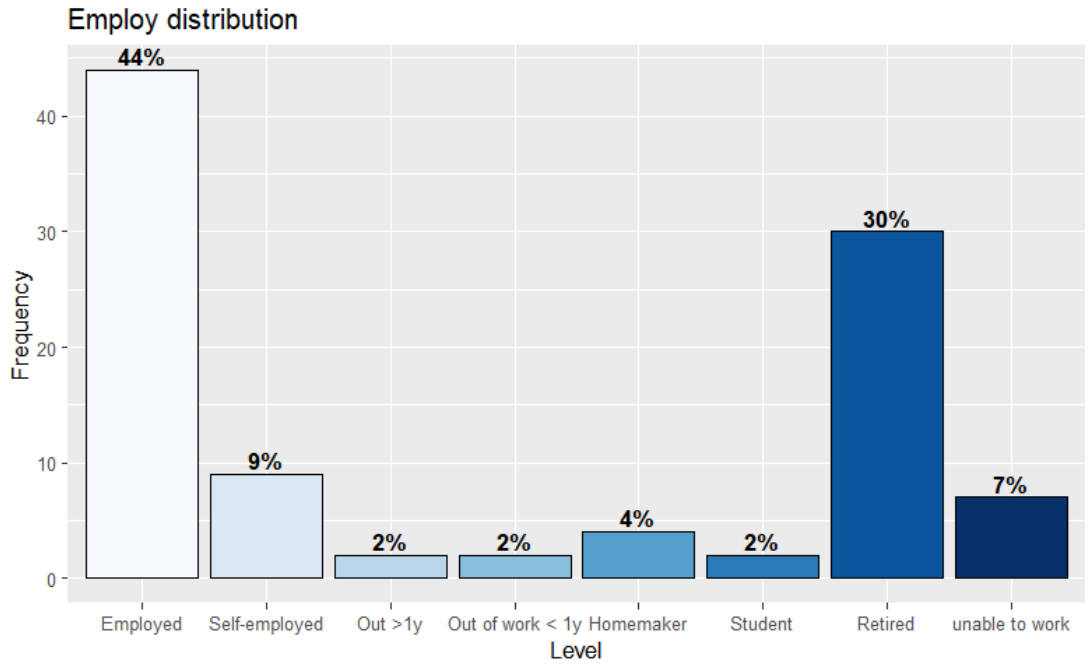

Fig. 3.5: Eight types of employ, in relative frequency.

sample, where the question is "Are you currently . . . ?", where eight are the possible answers: 1 stands for employed for wage, level 2 for self-employed, level 3 for out of work from more than one year, level 4 for out of work from less than one year, level 5 for homemaker, level 6 for student, level 7 for Retired, and level 8 stands for unable to work. As it is possible to see from Figure 3.5 the level with the major frequency is the employed for wage, and the second one is the retired category due to the huge presence of old adult respondents. Also, Figure 3.6 proof this tendency, where young adults, from 18 to 64 years, employed respondent are the most one and from 45 to 64 years female worker count is higher than male count.

The number and the distribution of the observation about the ownership of health coverage thanks to the variable `HEALTH.COVERAGE` that describes the question "Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare or Indian Health Service?" is then analyzed. As it is possible to see from Figure 3.7, where 1 is equal to "yes" and 2 is equal to "no", for each age, both males and females, the number of people with healthcare is far greater than those without.

From the Figure 3.8 is possible to observe the distribution of health coverage among the states, where a state with a dark color refers to a state with a smaller number of respondents who own health coverage, such as Texas (around 80%). Instead, most states are represented by lighter colors, like Rhode Island (around $96,25\%$) and Minnesota (around $95,86\%$), where there is a higher number of respondents who own health

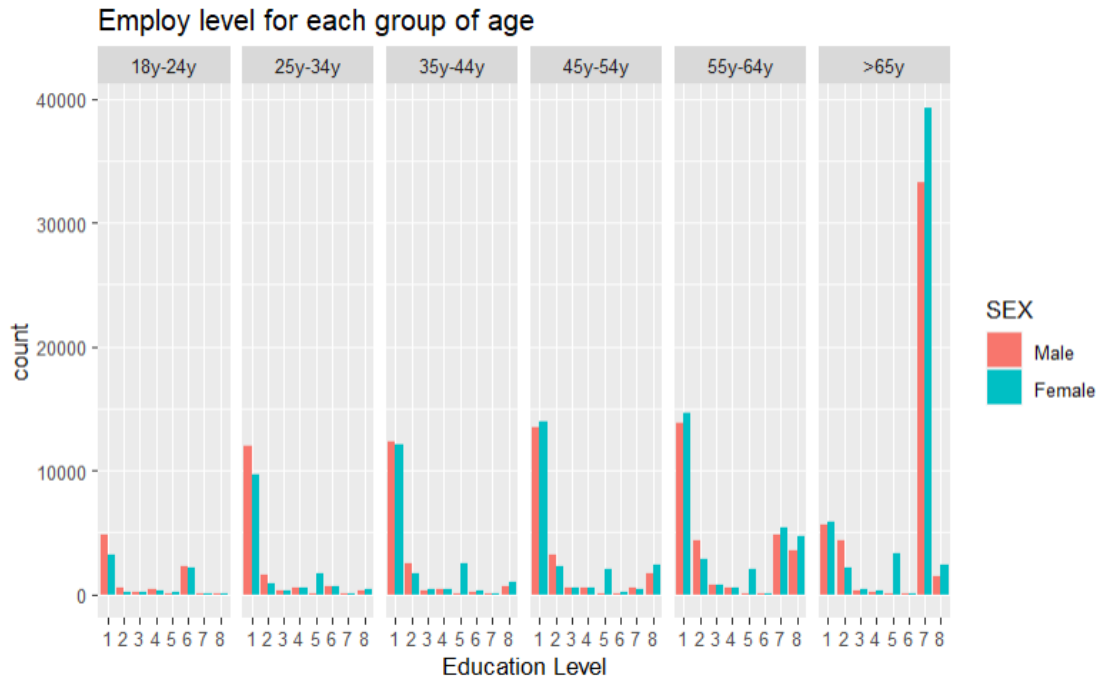

Fig. 3.6: Employ distribution per age and sex, where in the x-axis 1 stands for employed, level 2 for self-employed, level 3 for out of work from more than one year, level 4 for out of work from less than one year, level 5 for homemaker, level 6 for student, level 7 for Retired, and level 8 for unable to work.

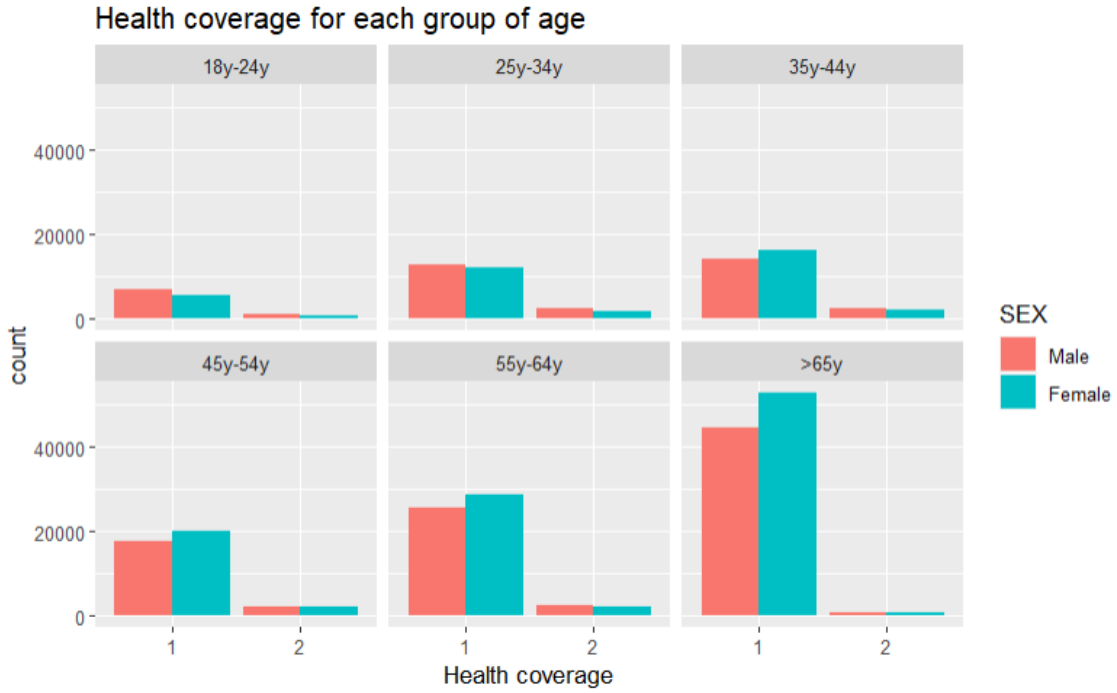

Fig. 3.7: Health coverage ownership in absolute frequency for each age group, where 1 stands for yes and 2 stands for no.

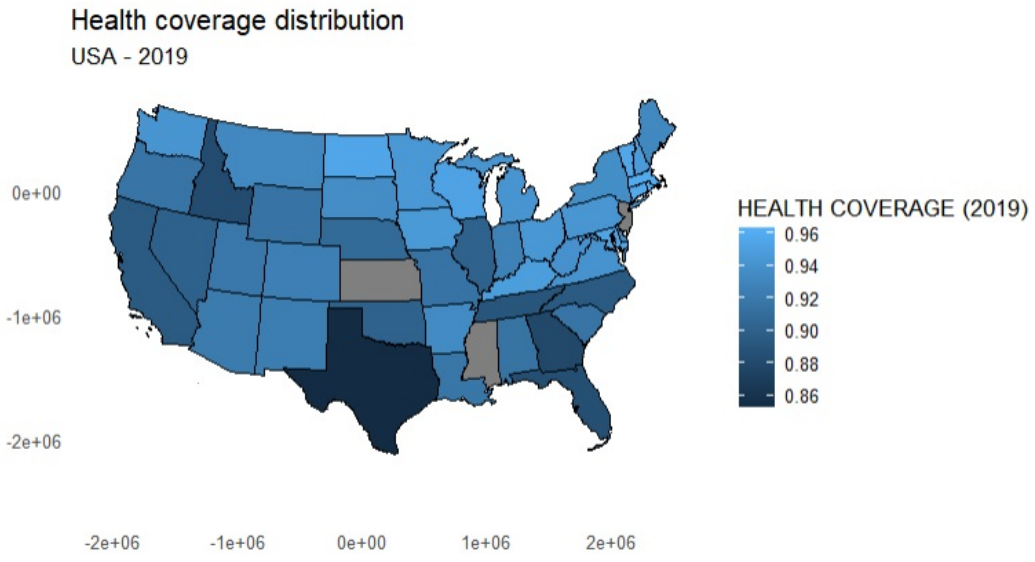

Fig. 3.8: Health coverage distribution for each state in relative frequency.

coverage.

3.2.2 Behavioural Risk Factor Analysis

This sub-section will analyze four behavioral risk factors for health: overweight and obesity, level of exercise, tobacco use, and alcohol use.

The first risk factor is alcohol use, to study this behavioral component, the variable `ALC.class`, a categorical variable created by the numeric column `AVEDRNK`, is analyzed. Thanks to the definition of binge drinking, in the project, the figure of the alcoholic has been outlined. Therefore, drinking refers to how much alcohol is consumed in a short period of time with the intent to feel drunk instantly. Each country has a different definition of binge drinking. In this case, it will be more than three drinks, both for men and women. The `AVEDRNK` measure is the answer to the following question: "One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. During the past 30 days, on the days when you drank, about how many drinks did you drink on average?". It refers to the average number of drinks taken at any time of the day when the citizen has decided to start drinking. From `AVEDRNK` is been created the categorical variable `ALC.class` divided into two levels: drinkers and not drinkers following the binge drinking concept. Figure 3.9 shows for each age class divided by sex, the absolute frequency of the `ALC.class` variable. We can note that men are more likely to get drunk than women, and in general, the age group that consumes more alcohol on average is people between 25 and 34 years old.

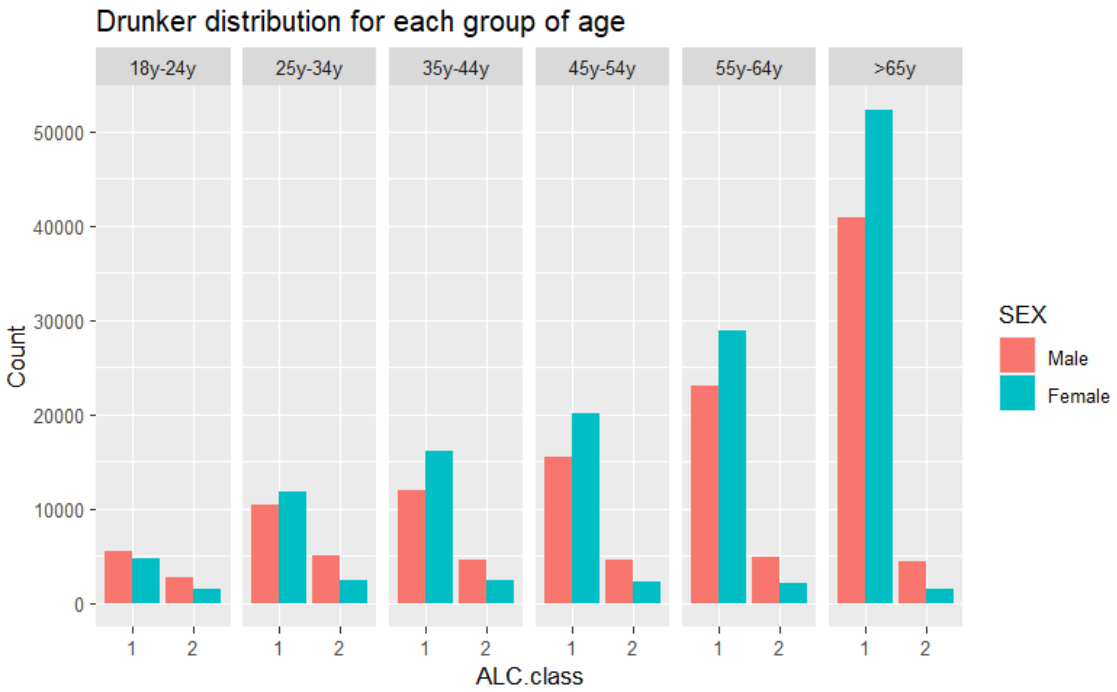

Fig. 3.9: Risk factor number 1: Alcohol use, where level 1 refers to not drinkers respondents and level 2 for drinkers, distribution of drunker per age and sex, absolute frequency.

The second risk factor is a behavioral component present especially in America that causes serious damage to the health of individuals: the overweight calculated by body mass index (`BMI`), defined as the ratio of respondent weight (in kilograms) by the respondent height (in meters). Four are the categories of `BMI` variable, and are the following:

- Underweight Respondents classified as underweight based on body mass index. ($\leq$ `_BMI5` < 18.50)
- Normal Weight Respondents classified as normal weight based on body mass index. ($18.50 <= \leq$ `_BMI5` < 25.00)
- Overweight Respondents classified as overweight based on body mass index. ($25.00 \leq$ `_BMI5` < 30.00)
- Obese Respondents classified as obese based on body mass index. ($30.00 <= \leq$ `_BMI5` < 99.99)

To simplify the study, the variable is been transformed from four levels to three levels by merging the third and fourth levels in order to consider the new third level as the group at risk. From Figure 3.10 it is possible to notice that more than 70% of respondents suffer from obesity or overweight in roughly similar parts for men (36%) and women (34%).

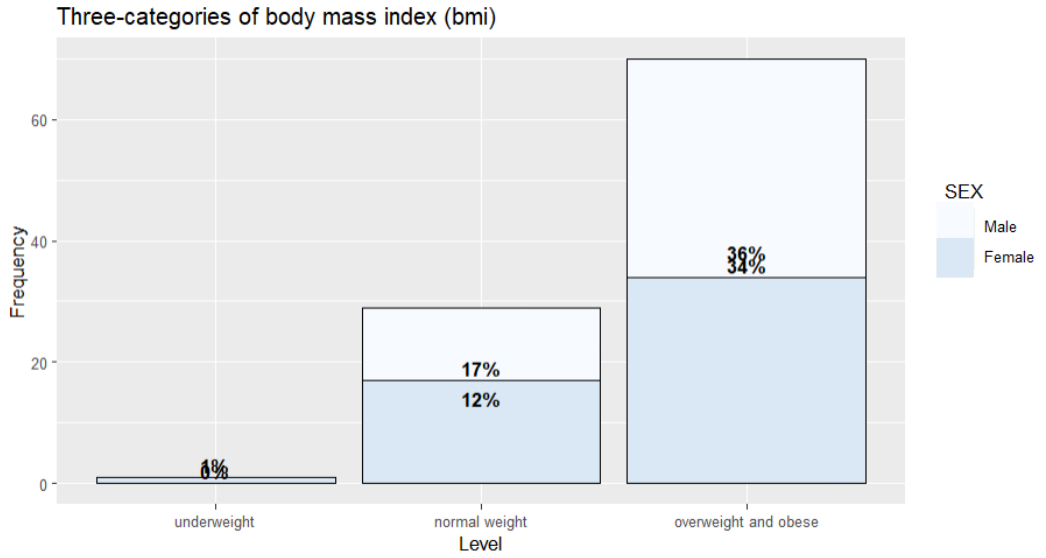

Fig. 3.10: Risk factor number 2: `BMI` relative frequency, where the percentage of underweight respondents is about 1% and overweight and obese respondents is about 70%.

The third risk factor is the smoke use. The questionnaire demand of `SMOKDAY` is "Do you now smoke cigarettes every day, some days, or not at all?", where the three possible answers were transformed into two: smoker and not.

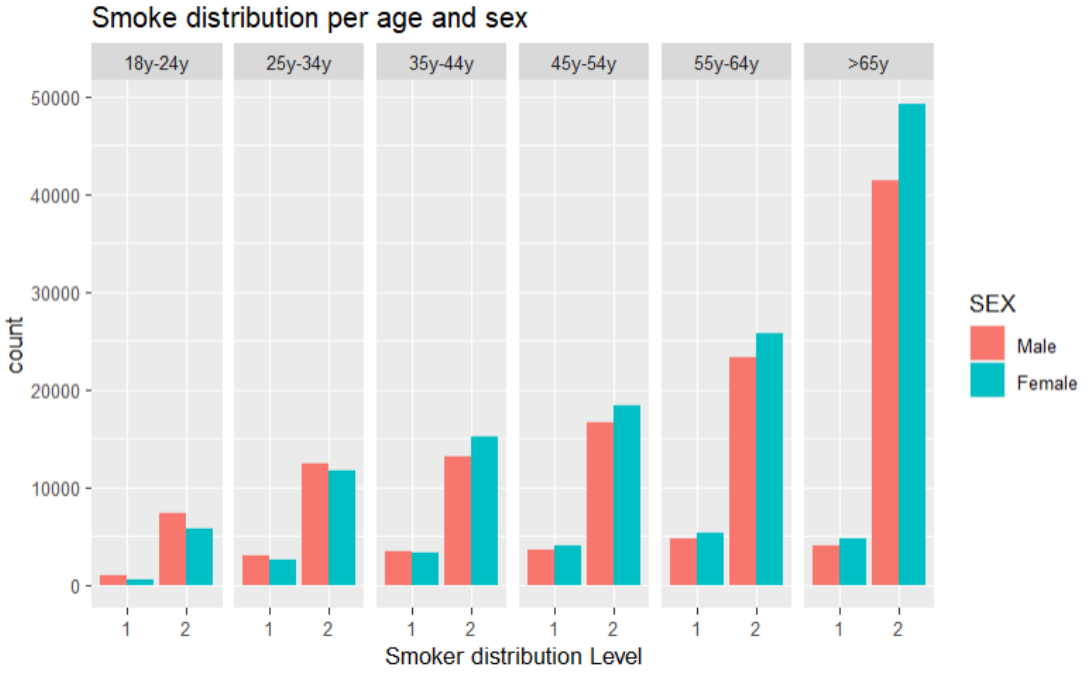

Fig. 3.11: Risk factor number 3: smoke use where level 1 refers to smokers and level 2 to respondents who don't smoke, in absolute frequency.

The three-level variable has been a two-level variable because it is decided to consider a smoker who answers every day and some days and a non-smoker who doesn't smoke at all. In Figure 3.11 it is possible to observe that smokers are more men than women at young ages. Instead, from 55 years old and upwards, women are more likely to smoke than men.

The last risk factor is the level of exercise for each respondent. The variable used is the categorical variable `EXERCISE`, that answer to the question is "During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?". The answer takes two levels: 1 means "yes" and 2 means "no". We will assume that people who answer yes because they have practiced regular physical activity in the past month are used to practicing sports constantly. In Figure 3.12 shows that a huge percentage of respondents for each age practice sport with regularity.

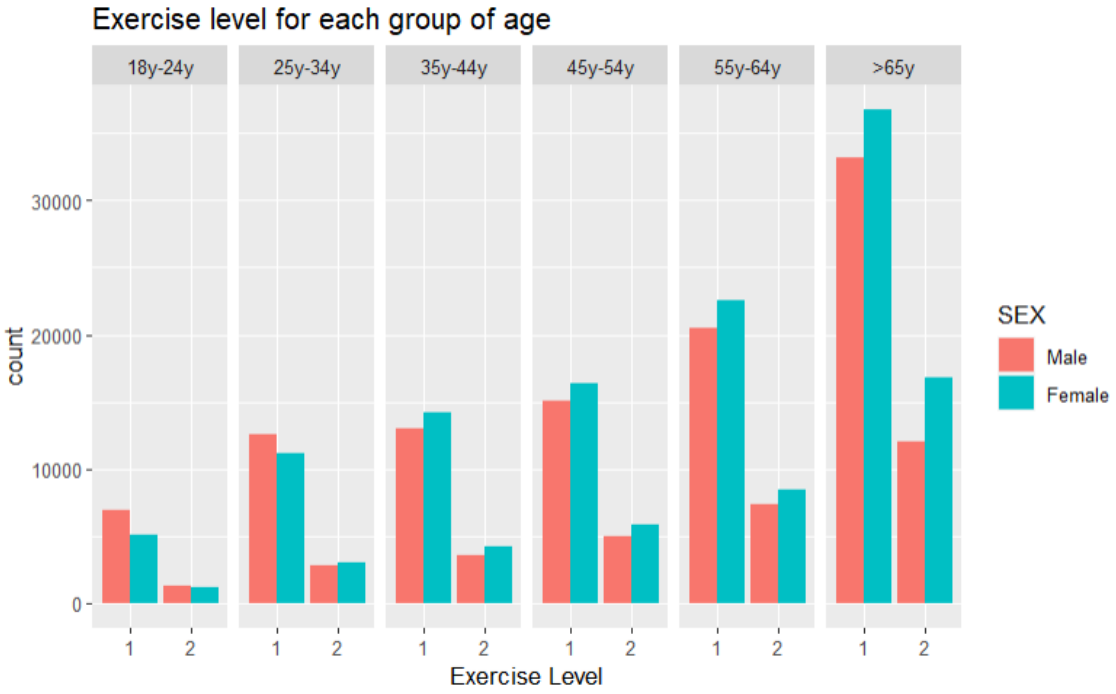

Fig. 3.12: Risk factor number 4: Physical activity, where level 1 refers to respondents who practice regular physical activity and level 2 refers to those who do not, distribution per age and sex in absolute frequency.

To analyze the trend of all four risk factors together is been create a dummy variable called `Risk_Factor`, where 1 code means the participant suffers at least from two risk factors, and 0 code means the respondent suffers from one or zero risk factor.

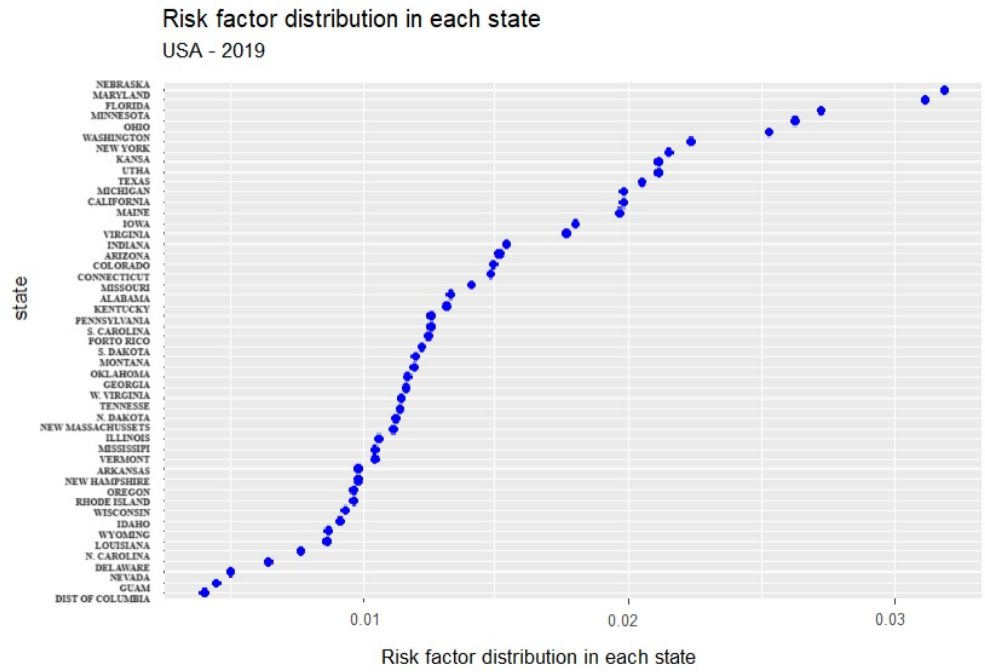

Fig. 3.13: Risk factor distribution for each state, the x-axis values are relative frequency.

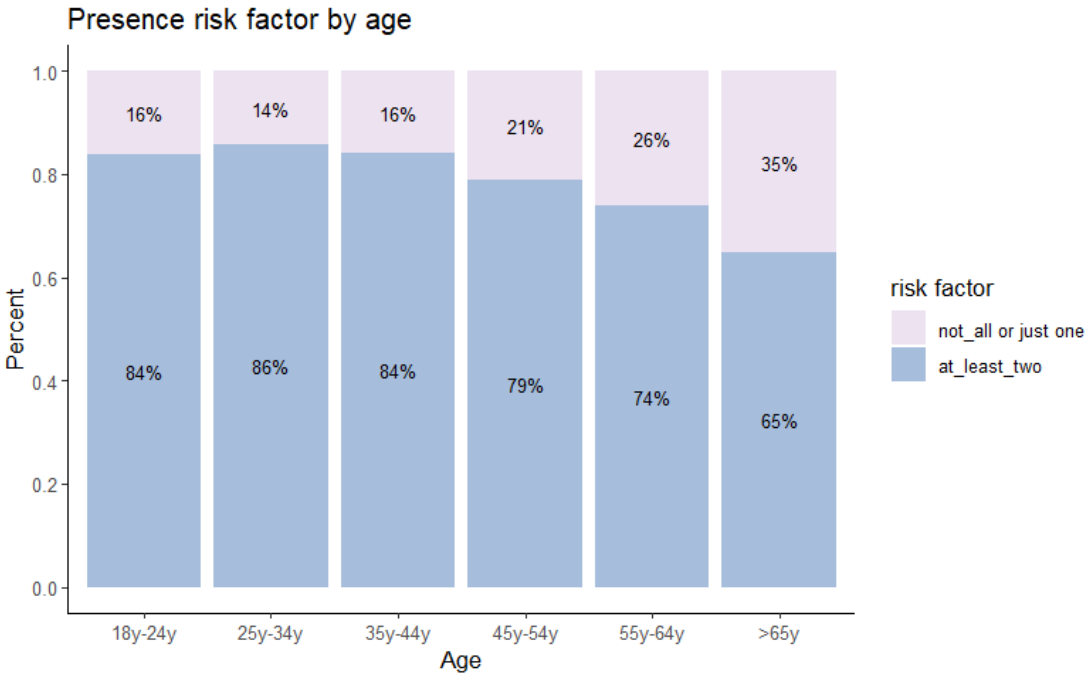

Fig. 3.14: Risk distribution per age.

From Figure 3.13 it is possible to observe that in Nebraska and Maryland, with a percentage higher than 3%, there is a major concentration of people who suffer from more than two risk factors previously listed. From Figure 3.14 it is possible to observe that the age class more affected by risk factors is the first one, where people between 18

and 24 years old are affected by risk per 84%, the second class between 25 and 34 years old are about 86% and the third one 84%. The oldest class, while suffering less from the presence of risk, still has a high percentage of risk (65%).

3.2.3 Cancer Disease Analysis

In this thesis, we decided to analyze the cancer disease (i.e., CANCER variable), that answer to the question "Ever told) (you had) any other types of cancer?" and the possible answers are: 1 equal to "Yes" and 2 equal to "No". Of our respondents in 2019, the percentage of people who have suffered from cancer is 10%.

A graph to investigate the variable CANCER is the bar plot in Figure 3.15, where is shown the distribution per AGE and SEX of the respondents who has or had cancer. The percentage of female (SEX =2) respondents with cancer is higher than the male one for each group of age from 25 years old. The last age bracket has the highest percentage of cancer disease, instead of the younger group has the lowest.

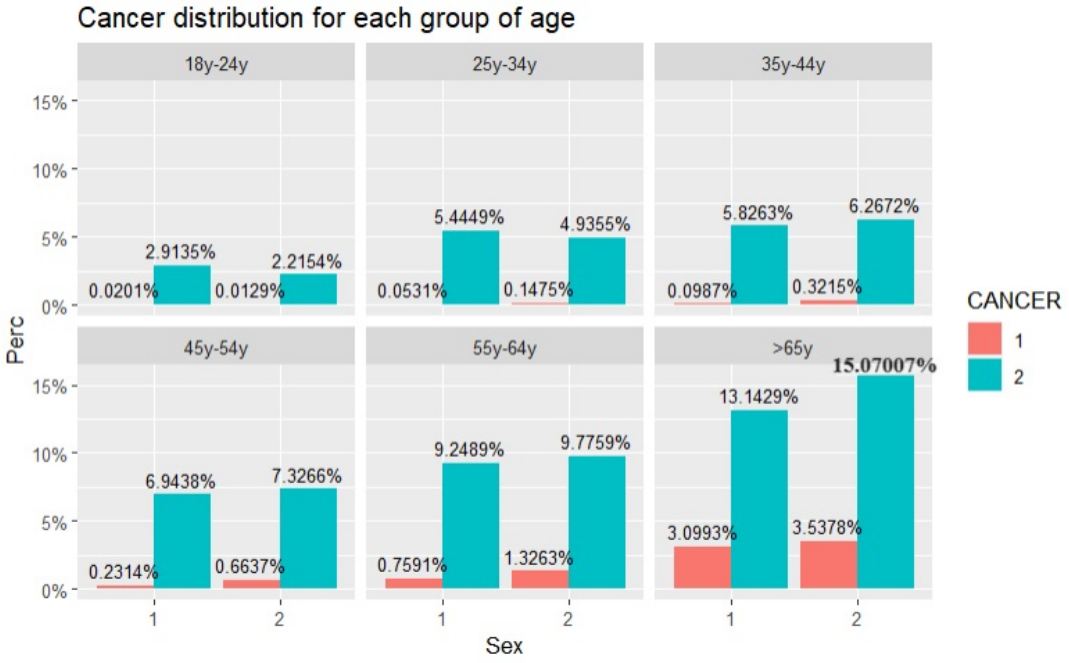

Fig. 3.15: Cancer distribution relative frequency per age and sex, where 1 ion x-axis means male and 2 means female, in the legend 1 means the observation have or had cancer and 2 means not.

To investigate the spread of the cancer around the USA territories, Figure 3.16 shows in which states with lighter color have the higher number of respondents who had or have suffered from cancer cases, such as Maine (around 13, 11%), Arkansas (around12, 46%),

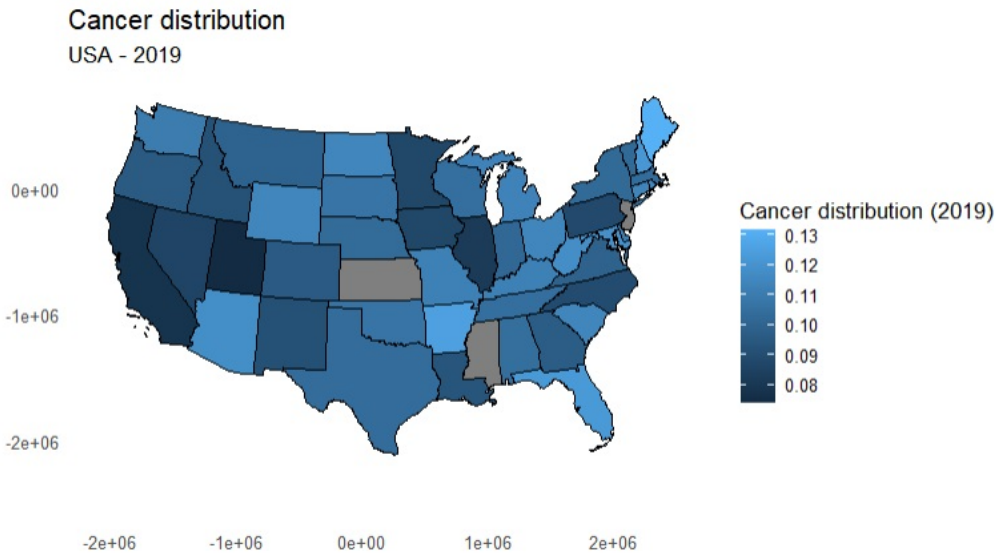

Fig. 3.16: Cancer distribution among the state, in relative frequency.

and Florida (around $12,20\%$). Instead, the states represented with the darker color have the lower number of respondents who have or had suffered from cancer, such as Utah (around $7,38\%$) and California (around $7,80\%$).

3.3 Spatial Clustering

"I invoke the first law of geography: everything is related to everything else, but near things are more related than distant things." [23]

In spatial clustering, the observations in a specific location are correlated with the observations in the neighboring locations. In health data science, data analysis is used to search for and identify groups, make diagnoses of clinical pictures, and predict cases of morbidity of individuals and populations. Cluster analysis permits reducing the data complexity and highlighting sub-populations characterized by different levels of socio-demographic variables, for example. In particular, clustering, such as k-means, uses distance-based measures, and the goal is to group observations according to their similarity from each other considering specific variables. To achieve this result, the unsupervised learning groups the observations of the dataset by unifying similar objects and separating dissimilar objects. Clustering is particularly useful when working with high-dimensional data, that is, data characterized by a huge number of variables.

However, in some data, spatial information, such as the BRFSS data, is available that traditional clustering methods do not take into account during the calculation process of the clusters. In this framework, spatial clustering allows researchers to examine whether certain risk factors, demographic data, health data, and many others tend to

cluster together, taking into account the specific geographic area information. This type of study can provide valuable insights for public health interventions and can address policy decision-making. By identifying (spatial) clusters of risk factors, researchers can better understand the underlying determinants of these health outcomes and develop targeted interventions to address them. The analysis typically involves geospatial data, such as demographic information, environmental factors, socioeconomic indicators, and disease prevalence data. Overall, (spatial) clustering of risk factors provides a valuable tool for understanding the spatial distribution of health risks, identifying high-risk areas, and informing targeted interventions and policies to improve public health outcomes.

In this thesis, we use the `skater` algorithm [22] that permits to apply a particular spatial clustering on the BRFSS data based on the graph theory explained in the next subsection. This approach enables to consider the spatial information of our observations (i.e., USA states) during the clustering process since socio-demographic variables, risk factors and health-related variables are spatially correlated.

3.3.1 Graph theory, Minimum Spanning Tree and Skater algorithm

The graph theory [24] investigates graphs, a mathematical object composed of nodes and edges that link these nodes. The graph theory is mainly used to study the structure of a network, to optimize problems where the aim is to find the shortest possible connection/edge path, in the data structure to represent the relationships between objects and entities, and in algorithm design to solve various computational problems.

In the context of graph theory, a Minimum Spanning Tree (MST) is a connected subgraph where nodes are connected by edges, minimizing the total edge weight. In this case, the edges are the states. Each state is represented as a point in the middle of its geographical area (i.e., centroid), the weights of the edges are instead the sum of the Euclidean distance between nodes considering the risk, demographic, and health variables that were previously analyzed. Therefore, the dataset used as the weight for the edge is composed of eleven variables transformed as percentages: `prop.smoker` represent the percentage for each state of smoker respondents, `prop.alchol` represent the percentage for each state of the drunker respondents, `prop.bmi` represent the percentage for each state of the respondents with obese or overweight problems, `prop.no.exerc` represent the percentage for each state of the respondents who don't practice regular physical activity, `prop.sex.f` represent the percentage for each state of female respondents, `prop.H.income` represent the percentage for each state of the respondents who earn an income higher than 65.000 dollars per year, `prop.old` represent the percentage for each state of respondents older than 45, `prop.young` represent the percentage for each state of respondents younger than 45, `prop.HC`

represent the percentage for each state of the respondents who own health coverage, `prop.edu` represent the percentage for each state of the respondents that attended college or have a college degree, `prop.cancer` represent the percentage for each state of the respondents who had or have cancer. This variable will be scaled before the use.

The MST is used to identify clusters by an algorithm called `skater` created by Assunção and coauthors in 2006 [22]. The algorithm prunes the corresponding MST to compute the clusters, pre-specifying the number of clusters (i.e., the number of MST subgraphs). Therefore, first of all, a graph must be created that describes the spatial structure of our data. So, a dissimilarity matrix is created, which represents the weights of the graphs. Then, the graph is converted into a minimum spanning tree, where the objects are connected with the minimum overall between-node dissimilarity. Finally, the MST is pruned, considering a pre-specifying number of clusters.

In this thesis, the problem of defining the number of clusters to be used is solved by the silhouette information [25]. Silhouette is one of the different methods to interpret and validate if a cluster algorithm is consistent. The silhouette value quantifies the degree of similarity between an object and its assigned cluster, reflecting the cohesion within the cluster and the separation from other clusters.The score ranges between -1 and 1, where a score of 1 means the specific data belongs to the cluster and is far from others, a -1 score means the worst value, and 0 represents an overlapping status. Equation (3.1) shows the formula to compute the silhouette $s(i)$ value for each data point i in cluster C_j, where $j \in \{1, \ldots, J\}$ and J is the total number of clusters considered. The $a(i)$ value measures the similarity of point i with the other point in the cluster, and $b(i)$ is the measure of dissimilarity of point i from the point in the other clusters. The dissimilarity is calculated using different kinds of distance metrics, as, for example, the Euclidean distance.

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}, \quad \text{if} \quad |C_j| > 1 \tag{3.1}$$

In order to have a silhouette measure for each cluster C_j, the mean of $s(i)$ where $i \in C_j$ is considered. Therefore, to have a global measure of the goodness of clustering considering a specific number of groups, the overall average of the $s(i)$ values is calculated.

In `R` code, the silhouette coefficient is calculated for each number of cuts of the dataset desired, from which the number of clusters is derived. The value of the optimal cut corresponds to the one that has the higher silhouette score, and the optimal number of clusters is one more than the number of cuts.

In addition to the Silhouette index, there are other methods: Intra-Cluster Correlation Coefficient (ICC) [26], Elbow Method, External Validity Index, and Internal Validity

Index [27].

3.3.2 R Code Analysis

To do the cluster analysis, a new dataset is required, which is derived by merging the USA `shapefile` dataset and a dataset derived from "BRFSS.ASCII" data composed of twelve variables. The eleven variables represent the percentage of the eleven factors described below that will scaled and used to weight the edge of the MST. The eleven rates used for each state are the following: rate of smokers, rate of drinkers, rate of overweight and obese people, rate of people who practice activity with regularity, rate of females, rate of people with high income, rate of young and old adults, rate of ownership of health coverage, rate of people with at least a high school degree, and rate of cancer cases. All these rates will be scaled before the use. Finally, the variable `state` was used to do the merge with the USA `shapefile`. The new merged dataset is now called `spatialData`, this data will again be divided into two parts, where `db_scaled` is the dataset that contains the scaled variable and `as.Spatial(spatialData)` is the `shapefile` data that contains the `geometry` variable that describes the initial graph structure.

The following code explains the construction of the minimum spanning tree and the use of the `skater` algorithm thanks to the `R` package called `spdep` [28]:

```
#Graph object
ct_nb_snap <- poly2nb(as_Spatial(spatialData))
costs_snap <- nbcosts(ct_nb_snap, data = db_scaled[,-1])
ct_w_snap <- nb2listw(ct_nb_snap, costs_snap, style="B")
#Minimum spanning tree
ct_mst_snap <- mstree(ct_w_snap)
#Contiguity-constrained cluster
clusj_snap <- spdep::skater(edges = ct_mst_snap[,1:2],
                            data = db_scaled[,-1],
                            ncuts = j)
```

At first, the neighbor list object is created using the function `poly2nb` on the `shapefile` data (i.e., `as.Spatial(spatialData)` object). Then, the function `nbcost` creates the edge cost weighted by using the scaled variable and excluding the `state` variable. Therefore, we use the function `nb2listw` to transform the neighbor list to a list weights object by using the just created `costs.snap` object. We consider the style "B" to not have the row-standardized, which means that the matrix of spatial weights is binary, where the value 1 indicates that two spatial units are contiguous and the value 0 that they are not, where two spatial units are contiguous means two geographic areas with at least one border in common. Then, the weights object is used in the `mstree` function

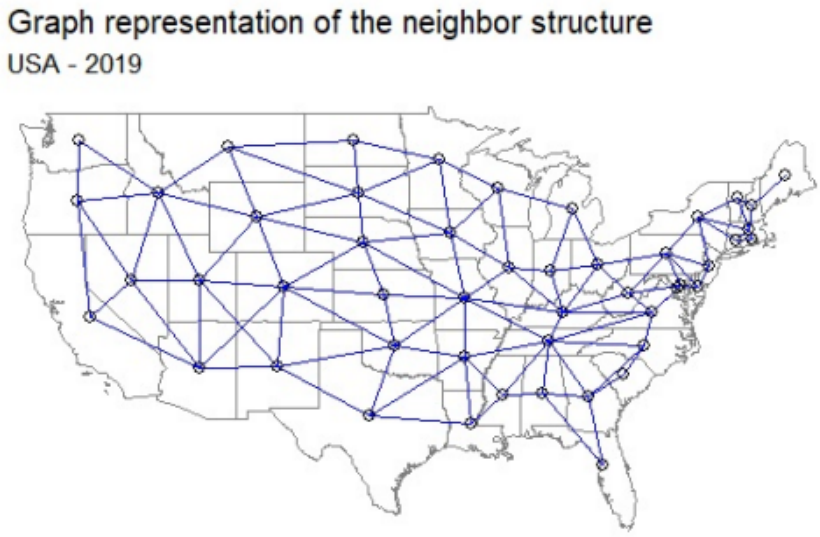

Fig. 3.17: USA initial graph with the edge and the centroids, used to do the spatial clustering analysis.

to create the minimum spanning tree. Lastly, the `skater` algorithm is applied to compute the clusters. The mandatory arguments of the algorithm are: the minimum spanning tree matrix column one and two `ct_mst_snap[,1:2]` (where the first column correspond to the starting node and the second column correspond to the ending node), the number of cuts (that are one less than the cluster number) `j` and the scaled data `db_scaled[,-1]`, excluding the `state` variable.

As said before, to create an MST is necessary to draw a connectivity graph that highlight the interconnection among object by using neighborhood associations. Each edge between spatial objects has a cost: the lower is the cost higher is the similarity between the states/nodes the edge joins. Once the weighted graph is constructed, it is time to create a minimum spanning tree, removing the edge with the higher cost. In the end, the graph is a connected graph with the minimum possible total edge weight [22]. Figure 3.17 shows the result of the initial connectivity graph created, where the centroids are represented by a little circle in the center of each state, and the edges are the blue lines that link the centroids. All the edges will be weighted by the sum of the Euclidean distance of the scaled variable Instead, Figure 3.18 represents the corresponding MST graph. The MST, by connecting all nodes in a graph with the minimum total cost, also reduces computational time by obtaining a simpler representation of the original problem in order to use fewer computational resources.

To decide the optimal number of clusters, the silhouette method is computed considering $\texttt{j} \in \{2, \ldots, 10\}$. The result is visible in Figure 3.19, that shows the `ncut` value, the higher silhouette score is when the number of cuts is equal to 2 represented in

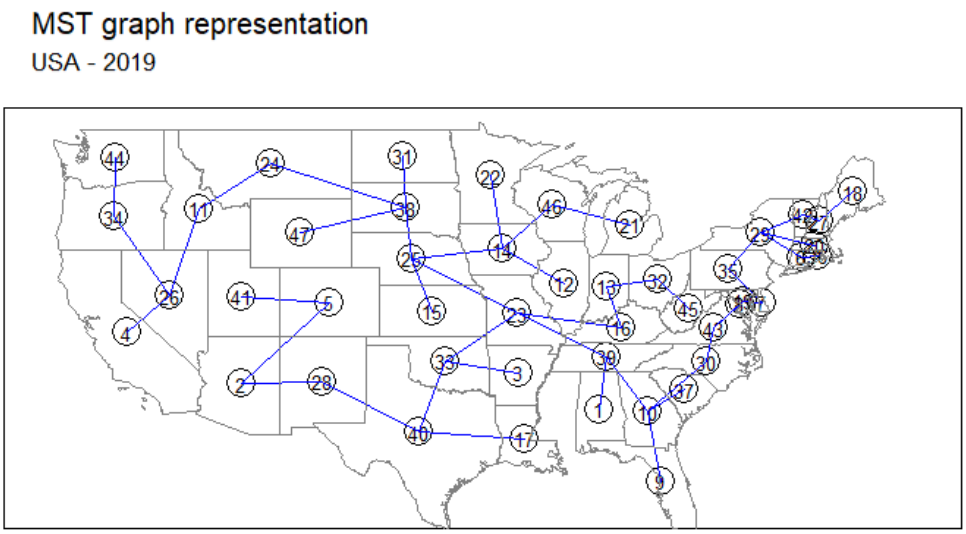

Fig. 3.18: USA MST graph representation.

Figure 3.19 by a red circle. The cut number is the times the dataset is divided (one less the number of the cluster), and three is the best number of clusters to improve the consistency of the analysis.

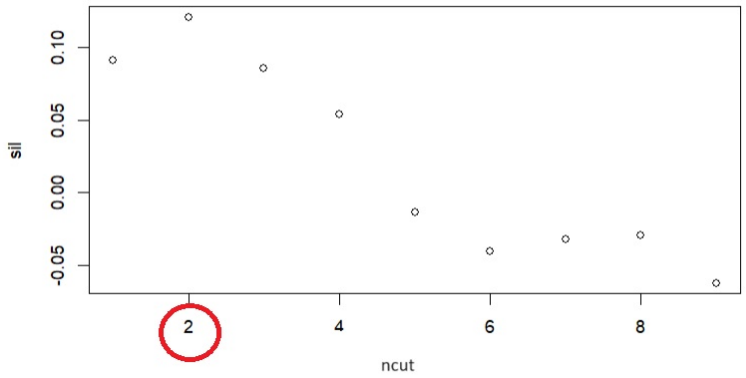

Fig. 3.19: Silhouette plot where the y-axis is the silhouette score, the x-axis is the number of cuts considered, i.e., the times the dataset is cut, where considering 2 cuts means dividing the dataset into 3 clusters.

Thanks to `skater` function, we finally apply the spatial constrained clustering approach in the BRFSS dataset. As said before, the algorithm takes into account the spatial contiguity in creating the clusters of states. Considering the optimal number of clusters, which equals 3, we find in the first cluster 19 states (Utah, Arizona, Colorado, New Mexico, Texas, Oklahoma, Missouri, Arkansas, Louisiana, Alabama, South Carolina, Tennessee, Kentucky, Indiana, Ohio, West Virginia, Georgia, Florida, North Carolina), in the second one 12 states (Maine, New York, Pennsylvania, Vermont, New Hampshire, Massachusetts, Rhode Island, Connecticut, New Jersey, Delaware,

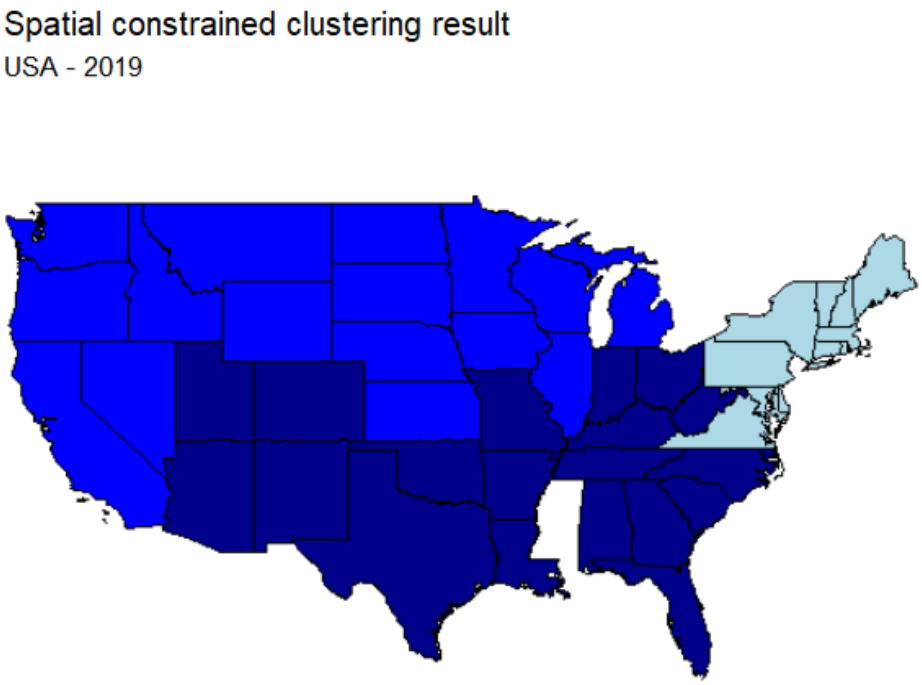

Fig. 3.20: The plot shows the result of the clustering analyses, where the three different colors refer to three relative clusters.).

Maryland, Virginia), and in the last one 16 (Washington, Oregon, California, Nevada, Idaho, Montana, Wyoming, Kansas, North Dakota, South Dakota, Nebraska, Illinois, Wisconsin, Michigan, Iowa, Minnesota). Figure 3.20 shows the spatial position of the clusters, identifying each cluster with a different color.

In addition, Figure 3.21 shows eleven box plots, one for each scaled variable used to weight the edge of the MST, from which is possible to observe the characteristics of each cluster. The first cluster is the most likely to smoke use and to not practice physical activity with regularity.The value of those who do not smoke and those who do not engage in physical activity is higher than the national average. The cluster also have obesity and overweight problems. Moreover, the number of females is lower than the national average. The positive aspect is the strong presence of young respondents and the very few cases of cancer. The second cluster is characterized by the higher number of respondents who smoke and do not practice physical activity regularly, earn a lower income per year, and a lower level of education, and have problems with being overweight and obese. The value of those who not practice physical activity, who smoke and who has weight problems is higher than the national average. At the same time, they have the fewer number of drinkers respondents, the fewer income level than the national average, the lower educational level, and the lower ownership of health coverage.

The third cluster, characterized by old people (the older than than the national

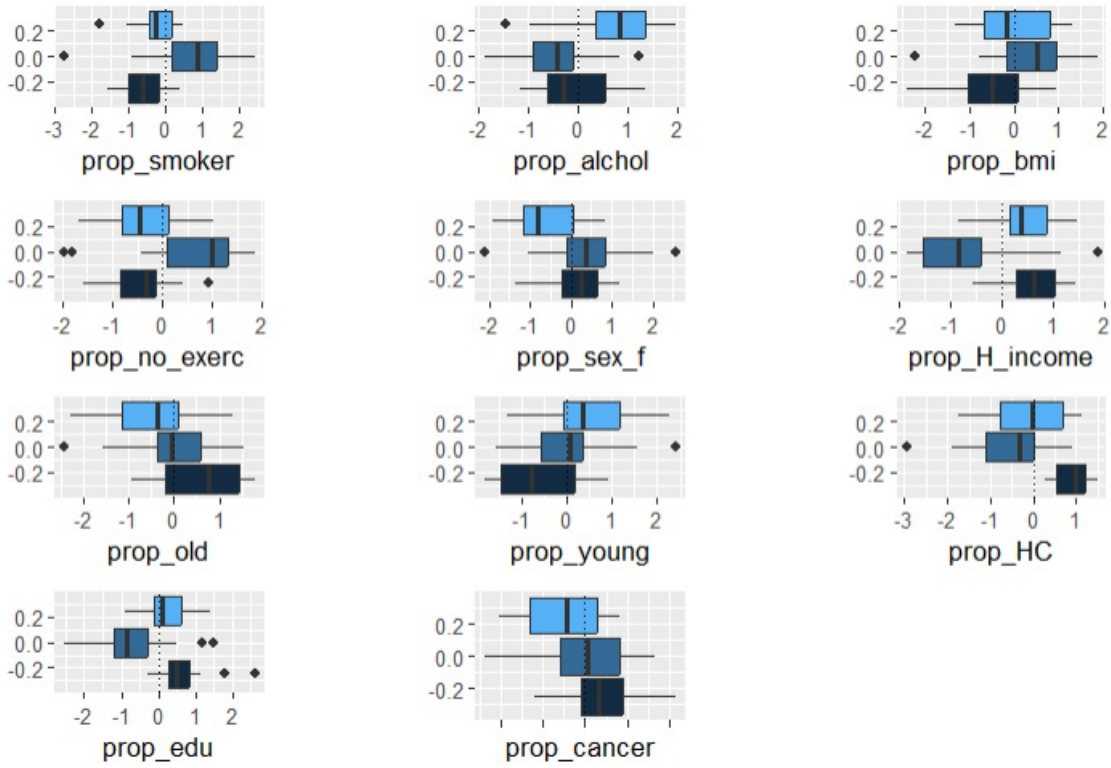

Fig. 3.21: Box-plot of 11 scaled variables for each cluster defined in Figure 3.20.

average), seems to be the healthier one because few people smoke, they have few problems of overweight, respondents practice exercise with regularity, and receive a high income. Moreover the level of education, the level of income per year and the level of the ownership of health coverage is higher than the national average. The only negative aspect is the higher number of respondents that suffer from cancer, the level is the higher than the national average.

In conclusion, the cluster with major risk factors is the cluster with more problems of income, less health coverage number, and lower education level. These results support the view that if the standard of living in a country is relatively poor, the presence of risk factors increases. The unhealthy lifestyle is evidenced by low education and below-average wages, which do not allow the population to take care of themselves. Regarding the variable cancer, we can see that a higher frequency is in the third cluster from a low presence of risk factors but, at the same time, populated only by people older than 45 years. It can be inferred that the factor that carries the most risk for the disease is age, as the respondents in that cluster are not heavy smokers, exercise regularly, and do not have weight problems. Differently, the other two clusters suffer more from the risk factors, but the distribution of the cancer variable indicates a lower occurrence of the disease in the two groups.

CHAPTER 4

Unveiling the Path to Healthier Communities: The Future of BRFSS and Health Surveys

This chapter will present a comprehensive overview of the findings of the analyses conducted on the BRFSS data and the interpretation of the result in the context of the research question and risk factor literacy. The discussion of the possible practical implications of the results and their potential impact on public health will be delineated. The value and potential opportunities, the challenges faced, and the future directions of big data analytics in healthcare will be discussed, trying to answer potential questions arising during the course of the analysis. To conclude the thesis, a section will discuss the performance of the spatial clustering analysis and the result derived from the algorithm, trying to give a solution for real-world issues through targeted health policy intervention.

4.1 Values, Challenges, and Future Directions of Health Survey, Data Surveillance

In healthcare, health surveys and data surveillance help in public health status monitoring by tracking public health trends [29]. All the data analyzed and studied from these methods can be useful for policymakers, healthcare providers, and researchers to

help in decision-making and increase the efficiency of the healthcare system. Thanks to the collection of data surveys is possible to identify health disparities among sub-populations characterized by different socio-economic levels, risk factors, and disease prevalence, preparing interventions to heal these gaps, using targeted interventions.

The challenges for these procedures regard different aspects: among these, there is the quality of the data due to the self-reporting method that may create bias and error, not ensuring reliability and completeness. Also, the privacy of the data is an important issue where the system needs to protect sensitive personal health information. In this case, BRFSS provides to take into account the privacy and the confidentiality of the questionnaire respondents, thanks to the anonymization of the data. Another important challenge is the participant engagement and response rates. In the BRFSS of 2019, only $200,000$ people, the 0.006%, have decided to compile the questionnaire against the $329,256,465$ USA residents over 18 years old.

The future directions of health surveys are subject to ongoing research and dynamics in health care. They could regard the use of digital technologies to collect data, such as mobile apps and wearable devices, that could provide data continuously and updated with less expenditure of resources. Another challenge might be to find a way to merge and analyze data from different sources, seeking a universal language for health data. Thanks to the use of innovative analytics technique [30], it is more and more immediate to identify new trends and patterns with machine learning and AI algorithms, such as the `skater` algorithm used in this thesis. The final scope of the poll is both that community involvement in the programs can increase awareness of self-care and lead participants to increase their physical health and that public government focus on health disparities and inequalities to create targeted interventions and policies to reduce disparities.

4.2 Exploring the Use of Big Data in Healthcare Analysis

Not all physicians support the use of analytical methods such as health surveys and data surveillance in healthcare, especially due to the fear that the privacy of the data will not be respected. For the sake of parity, it was decided to include this short section in which I try to give an answer to some of the different questions that arise about the use of big data in healthcare analysis [1].

The first demand is: Are individuals who participate in health surveys or health data surveillance aware of the consequence of giving personal data? Does data protection in the health area really protect the privacy of the data? The analysis of big data could have a significant impact on daily life, which could cause negative effects, like the leakage of private information and the denial of health insurance. Not all people

are aware of the risk they face due to the disclosure of personal data, either through misinformation or the complex nature of this data. To fight against this obstacle is necessary to implement data protection practices, inform about privacy rights and regulations, and support policies that prioritize privacy protection. Privacy in healthcare has a growing recognition. For this reason, several laws have been enacted to safeguard individuals' data, such as the European General Data Protection Regulation (GDPR) and the American Health Insurance Portability and Accountability Act (HIPAA). The balance between big data's positive effects and data privacy is faltering and is a continuous source of discussion. According to Lyon, author of "Surveillance, Snowden, and Big Data: Capacities, consequences, critique" [31] for example, the importance of technology, data collection through data surveillance, and the ability to analyze massive amounts of data are not questioned, but the reliability of these procedures, the quality of data and the methods are. The article discusses the need to associate ethical considerations with the analyses in order to have a better guarantee of privacy, which is made vulnerable by this type of approach, suggesting that technical boundaries should be established and that the concept of privacy itself should be redefined in the context of big data.

The second question is: [32] After discovering, from the data analysis, a potential strategy to improve health conditions, how is the veracity of discovery certified? The result of a study usually leads to a pattern to follow to solve an initial problem, in our case, the spread of risk factors in USA states, but to assure the veracity of the pattern, different assurance mechanisms are used. One method could be the validation and replication of the study with an independent dataset to ensure the consistency of the patterns. Another way is to compare the information with the existing scientific knowledge to obtain an aligned conclusion. Also, a peer review is a solution, an analysis done by experts that corrects bias and error in data interpretation, that should be clear, reliable, and transparent to allow replication.

The third question is about the relationship between new analytics methods versus traditional clinic research, where data is typically collected through manual methods such as paper-based forms, interviews, and physical examinations, and their respective importance in the research area [33]. There is no unique answer to this question cause the interconnection between the two methods is multifaceted and complicated. Each of them has strengths and weaknesses: the first one allows the collection of a huge amount of data, identifying trends and correlation more easily than the second, being a valuable resource for disease prevention and health population management. The second, in turn, is most useful in the case of a specific research question based on a hypothesis-driven path thanks a carefully planned study designs and consistent data collection. Based on the type of research question, one method could be more suitable than the other. For example, analytics is more convenient for exploring large datasets and identifying associations, but

data protection is not guaranteed. Instead of traditional research is useful to establish causality but is expensive and time-consuming.

The last question is: How willing is the population really to learn about their health through findings derived from big data? Most importantly, how willing are citizens to share their health information? The willingness to know deeper the own health status depends on the character of an individual [34], for some, having more knowledge in the health field leads to a healthier and more informed lifestyle. The ability to receive this information in a direct and immediate way motivates individuals to healthier and more balanced behaviors, becoming active seekers of good health. At the same time, this information can create anxiety and preoccupation due to misinterpretation of data, such as the deleterious move when we look for the cause of symptoms on the internet. Always placing a greater emphasis on the fear of a possible invasion of privacy by unauthorized third parties. Factors such as education, health literacy, and other factors can also influence individuals' willingness and capacity to engage with health information derived from big data. From the demographic point of view, for example, younger individuals are more motivated to share health data for research purposes, even if in the BRFSS dataset, we noted a high rate of old people. Also, people with higher socioeconomic status are more motivated to share information, and also, in the project analysis, the major part of observations have a high income. This is because they have better access to healthcare resources and more time and money to spend on their health and to prioritize it. People with diseases also share data more openly with the hope of finding a cure sooner rather than later or better treatments. Different trends in sharing data based on socio-demographic and health factors lead to bias in the study. They bring bias into the research and risk making the samples unrepresentative, over-representing the sick, and making generalization in the research more difficult. It is therefore necessary to engage in data privacy so that more people join in releasing their data in order to increase the representation of the study.

4.3 Spatial Clustering and Personal Observation

The analysis of spatial clustering about risk, demographic, and health factors showed that it is essential to consider the spatial coordinates of a sub-population when investigating health-related issues. We could take into account the citizens of the state of the first cluster (Washington, Oregon, California, Nevada, Idaho, Montana, Wyoming, Kansas, North Dakota, South Dakota, Nebraska, Illinois, Wisconsin, Michigan, Iowa, and Minnesota), where from the box-plot represented in Figure 3.21 we notice that is very high the number of respondents who suffer from alcoholism, obesity, and overweight, as well as low practice of regular physical activity. Taking these elements into consideration, an

analysis can be done that understands whether there is any factor in that area that affects these behaviors, and a solution can be sought to resolve them. The different clusters brought to light the disparities due to the different communities that could be fixed by a public health government with interventions and resources specific to the situation. For example, for the citizens of the state of the second cluster (Maine, New York, Pennsylvania, Vermont, New Hampshire, Massachusetts, Rhode Island, Connecticut, New Jersey, Delaware, Maryland, and Virginia), the public government could invest in education and skill training to entice more children to study to get better paying jobs and consequently have the economic ability to have health insurance. Another solution could be the initiative to support labor-intensive sectors, a job in a sector that requires a larger workforce, by promoting their development. Besides investigating demographic factors such as age, sex, and income, spatial analysis can also treat chronic diseases and could help by giving prevention solutions or healthcare services. The disease studied in our case is cancer. From the analyses, it is possible to observe that the number of respondents of the state in the third clusters (Utah, Arizona, Colorado, New Mexico, Texas, Oklahoma, Missouri, Arkansas, Louisiana, Alabama, South Carolina, Tennessee, Kentucky, Indiana, Ohio, West Virginia, Georgia, Florida, North Carolina) is the highest respect the other clusters' respondents. It is also possible to notice that in the third cluster, there is the highest concentration of older people, a factor that could be linked to the high frequency of the disease. One solution to try to reduce cancer deaths could be to implement screening programs for early detection of certain types of cancer or to promote vaccination to prevent infection with viruses that can cause cancer. This research can be developed using a larger number of risk, demographic, and health factors, leading to a more comprehensive analysis, ultimately leading to more effective strategies.

Talking about the future directions of BRFSS there are several there are several possible branches. The most obvious ones may involve the study of new emerging risk factors, such as Coronavirus, a new highly contagious virus spread in 2020. Another risk factor could be air pollution, waste pollution, and water pollution, an aspect that is growing sharply over time. Moreover, collect more socio-demographic variables that may be optimal to define sub-populations in more detail, like a variable to describe the level of urbanization, the types of access to health services, or the type of work performed, more or less sedentary and the risks involved in a given job. Regarding our analysis, future directions could include variables that already exist but have only been mentioned in this project, such as variable that describes the physical and mental health of an individual (`PHYSHLTH`, `MENTHLTH`). An analysis could also be carried out by considering more clusters than three, although not optimal, but with the aim of investigating the distribution of the variables in the USA territory in more detail. A proper comparison of the clustering results could be made by performing the same study using clustering approaches without

the spatial constraint.

Personally, I think that big data, health surveys, data surveillance, and several other technological tools can improve public health. Through the monitoring of health data, it is possible to have a comprehensive overview of diseases and risk factors effects, enabling prevention and minimizing the negative impacts. While it is important to acknowledge the associated risks, the benefits of these advancements far outweigh them.

REFERENCES

[1] K. Marconi and H. Lehmann, *Big data and health analytics*. CRC Press, 2014.

[2] B. L. Ryan, J. B. Brown, R. H. Glazier, and B. Hutchison, "Examining primary healthcare performance through a triple aim lens," *Healthcare Policy*, vol. 11, no. 3, p. 19, 2016.

[3] R. M. Bohmer, A. EDMONDSON, and L. R. FELDMAN, "Intermountain health care," 2002.

[4] S. Galea, S. M. Abdalla, and J. L. Sturchio, "Social determinants of health, data science, and decision-making: Forging a transdisciplinary synthesis," *PLoS Medicine*, vol. 17, no. 6, p. e1003174, 2020.

[5] B. Steffen, "Social determinants of health," 2023.

[6] J. M. Soucie, "Public health surveillance and data collection: general principles and impact on hemophilia care," *Hematology*, vol. 17, no. sup1, pp. s144–s146, 2012.

[7] H. Ashrafian and A. Darzi, "Transforming health policy through machine learning," *PLoS Medicine*, vol. 15, no. 11, p. e1002692, 2018.

[8] M. I. Razzak, M. Imran, and G. Xu, "Big data analytics for preventive medicine," *Neural Computing and Applications*, vol. 32, pp. 4417–4451, 2020.

[9] M. A. Barrett, O. Humblet, R. A. Hiatt, and N. E. Adler, "Big data and disease prevention: from quantified self to quantified communities," *Big data*, vol. 1, no. 3, pp. 168–175, 2013.

[10] S. Artiga, E. Hinton, *et al.*, "Beyond health care: the role of social determinants in promoting health and health equity," *Kaiser Family Foundation*, vol. 10, 2018.

[11] C. J. Murray and A. D. Lopez, "Evidence-based health policy—lessons from the global burden of disease study," *Science*, vol. 274, no. 5288, pp. 740–743, 1996.

[12] C. J. Murray and A. D. Lopez, "Measuring the global burden of disease," *New England Journal of Medicine*, vol. 369, no. 5, pp. 448–457, 2013.

[13] World Health Organization and others, "Metrics: disability-adjusted life year (daly)," 2016.

[14] W. Giles, A. H. Mokdad, and D. F. Stroup, "Public health surveillance for behavioral risk factors in a changing environment: recommendations for the behavioral risk factor surveillance team," 2003.

[15] F. A. Batarseh, I. Ghassib, D. Chong, and P.-H. Su, "Preventive healthcare policies in the us: solutions for disease management using big data analytics," *Journal of big Data*, vol. 7, pp. 1–25, 2020.

[16] C. for Disease Control, Prevention, *et al.*, "Behavioral risk factor surveillance system (brfss)," *Atlanta: CDC*, 2014.

[17] M. Qayad, C. Pierannunzi, P. P. Chowdhury, S. Hu, G. M. Town, and L. Balluz, "Landline and cell phone response measures in behavioral risk factor surveillance system," *Survey Practice*, vol. 6, no. 3, 2013.

[18] C. Pierannunzi, S. S. Hu, and L. Balluz, "A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (brfss), 2004–2011," *BMC medical research methodology*, vol. 13, no. 1, pp. 1–14, 2013.

[19] M. I. Lapid, B. L. Clarke, and R. S. Wright, "Institutional review boards: What clinician researchers need to know," in *Mayo Clinic Proceedings*, vol. 94, pp. 515–525, Elsevier, 2019.

[20] V. Anant, L. Donchak, J. Kaplan, and H. Soller, "The consumer-data opportunity and the privacy imperative," *Retrieved from McKinsey website: https://www. mckinsey. com/business-functions/risk-and-resilience/our-insights/the-consumer-data-opportunity-and-the-privacy-imperative*, 2020.

[21] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[22] R. M. Assunção, M. C. Neves, G. Câmara, and C. da Costa Freitas, "Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees," *International Journal of Geographical Information Science*, vol. 20, no. 7, pp. 797–811, 2006.

[23] H. J. Miller, "Tobler's first law and spatial analysis," *Annals of the association of American geographers*, vol. 94, no. 2, pp. 284–289, 2004.

[24] O. Sporns, "Graph theory methods: applications in brain networks," *Dialogues in clinical neuroscience*, 2022.

[25] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[26] S. Killip, Z. Mahfoud, and K. Pearce, "What is an intracluster correlation coefficient? crucial concepts for primary care researchers," *The Annals of Family Medicine*, vol. 2, no. 3, pp. 204–208, 2004.

[27] S. Nawrin, M. R. Rahman, and S. Akhter, "Exploreing k-means with internal validity indexes for data clustering in traffic management system," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.

[28] Roger Bivand, "R packages for analyzing spatial data: A comparative case study with areal data," *Geographical Analysis*, vol. 54, no. 3, pp. 488–518, 2022.

[29] L. O. Gostin, L. A. Levit, S. J. Nass, *et al.*, "Beyond the hipaa privacy rule: enhancing privacy, improving health through research," 2009.

[30] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021.

[31] D. Lyon, "Surveillance, snowden, and big data: Capacities, consequences, critique," *Big data & society*, vol. 1, no. 2, p. 2053951714541861, 2014.

[32] J. L. Johnson, D. Adkins, and S. Chauvin, "A review of the quality indicators of rigor in qualitative research," *American journal of pharmaceutical education*, vol. 84, no. 1, 2020.

[33] B. Meskó, Z. Drobni, É. Bényei, B. Gergely, and Z. Győrffy, "Digital health is a cultural transformation of traditional healthcare," *Mhealth*, vol. 3, 2017.

[34] B. Auxier, L. Rainie, M. Anderson, A. Perrin, M. Kumar, and E. Turner, "Americans and privacy: Concerned, confused and feeling lack of control over their personal information," 2019.