



UNIVERSITAT DE
BARCELONA

Bachelor's degree
**COMPUTER ENGINEERING
DEGREE**

Faculty of Mathematics and Computer
Science
University of Barcelona

**DATA MINING AND VISUALIZATION OF
MULTI-SOURCE BIG DATA IN THE
UK-BIOBANK**

Author: Alex Domínguez Ortega

Director: Dr. Karim Lekadir
**Made in: Math and computer science
i sistemes informàtics**

Barcelona, June 20, 2021

Abstract

Big data has benefits and complications. Analysing big data has the capability of obtaining great knowledge which can be used for taking the best decisions, but it is an execution of high complexity.

In medicine, the amount of data generated every day is more than enough to be called big data. In this project, the data extracted from radiomics, an image analysis technique that generates about 650 different variables, is treated with the objective of being visualized in an understandable and intuitive way. These visualizations are compiled in the Radiomic wide association study web page, a tool that will help investigators formulating, confirming or refuting their theories, which could be used to support new health politics.

Resumen

El big data tiene beneficios y complicaciones. El análisis de big data tiene la capacidad de obtener un gran conocimiento que puede ser utilizado para tomar las mejores decisiones, pero es una ejecución de alta complejidad.

En medicina, la cantidad de datos que se generan cada día es más que suficiente para ser llamada big data. En este proyecto, los datos extraídos de las radiomics, una técnica de análisis de imágenes que genera alrededor de 650 variables diferentes, son tratados con el objetivo de ser visualizados de forma comprensible e intuitiva. Estas visualizaciones se recopilan en la página web Radiomic wide association study, una herramienta que ayudará a los investigadores a formular, confirmar o refutar sus teorías para soportar nuevas políticas sanitarias.

Resum

El big data té beneficis i complicacions. L'anàlisi de big data té la capacitat d'obtenir un gran coneixement que pot ser utilitzat per a prendre les millors decisions, però és una execució d'alta complexitat.

En medicina, la quantitat de dades que es generen cada dia és més que suficient per a ser anomenada big data. En aquest projecte, les dades extretes de les radiomics, una tècnica d'anàlisi d'imatges que genera al voltant de 650 variables diferents, són tractades amb l'objectiu de ser visualitzades de manera comprensible i intuïtiva. Aquestes visualitzacions es recopilen en la pàgina web Radiomic wide association study, una eina que ajudarà els investigadors a formular, confirmar o refutar les seves teories i suportar noves polítiques sanitàries.

Contents

1	Introduction	1
2	Objectives	3
3	Planning	5
4	Theoretical framework	7
4.1	Radiomics	7
4.2	P-values	9
5	Methodology	12
5.1	Dataset: UK-BioBank	12
5.2	Technologies used	13
5.3	Radiome wide association study web page	15
5.3.1	Homepage	15
5.3.2	Associations with developmental, lifestyle, clinical and environmental influences	16
5.3.3	Associations of radiomics with clinical outcomes	21
5.3.4	Association of radiomics with cardiovascular risk factors	24
5.4	Deployment	27
6	Use cases	29
7	Examples of use	33
8	Conclusion	53
9	Bibliography	54

List of Figures

1	Initial planning	5
2	Example of a cardio magnetic resonance and the designed regions of interest	7
3	3D renders of the left ventricle, right ventricle and left ventricle myocardium	8
4	First-order radiomic of the left ventricle myocardium	8
5	GLCM matrix and GLRLM matrix	9
6	Graphical representation of the flux of radiomics processation extracted from <i>Repeatability of Cardiac Magnetic Resonance Radiomic</i> [1]	9
7	Homepage of radiome wide association study web page	15
8	Data showcase section without selection	16
9	Data showcase section with the histogram	17
10	Help pop up of the Data showcase section	18
11	Side bar for the Odds section's field selection	18
12	Manhattan plot in the Odds section with selected fields	20
13	Information displayed on hoverign a dot in the manhattan plot	21
14	Help pop up of the Odds section	21
15	Side bar for the Correlations section's field selection	22
16	Heatmap in the Correlations section with selected fields	23
17	Help pop up of the Correlations section	23
18	Side bar for the Cardiovascular risk factors section's field selection	24
19	Chords plot of the Cardiovascular risk factors section for sex, smoking, BMI and shape radiomics feaetures of the right ventricle in diastole phase	26
20	Help pop up of the Cardiovascular risk factors section	26
21	Use cases for the Radiome wide association study web page	29
22	Correlations between shape radiomics features of the left ventricle in end of diastole phase and clinical outcomes	33
23	Correlations between shape radiomics features of the left ventricle in end of systole phase and clinical outcomes	34
24	Correlations between shape radiomics features of the right ventricle in end of diastole phase and clinical outcomes	35
25	Correlations between shape radiomics features of the myocardium in end of diastole phase and clinical outcomes	35
26	Volume of the left ventricle in diastole phase for males and females	36
27	Surface area of the left ventricle in diastole phase for males and females	36
28	Major axis of the left ventricle in diastole phase for males and females	37
29	Least axis of the left ventricle in diastole phase for males and females	37
30	Correlations between first-order radiomics features of the right ventricle in the end of diastole phase and clinical outcomes	38
31	Correlations between first-order radiomics features of the right ventricle in the end of systole phase and clinical outcomes	39

32	Entropy of the right ventricle in diastole phase for smokers and non smokers	39
33	Interquartile range of the right ventricle in diastole phase for smokers and non smokers	40
34	Mean absolute deviation of the right ventricle in diastole phase for smokers and non smokers	40
35	Volume of the left ventricle in diastole phase for smokers and non smokers	41
36	Volume of the right ventricle in diastole phase for smokers and non smokers	42
37	Volume of the left ventricle in diastole phase for patients older and younger than 60 years old	42
38	Uniformity of the left ventricle in diastole phase for patients with or without high tension	43
39	Volume of the left ventricle in diastole phase for males and females	44
40	Surface area of the left ventricle in diastole phase for males and females	44
41	Correlations between shape radiomics features of the left ventricle in diastole phase and clinical outcomes	45
42	Correlations between first-order radiomics features of the left ventricle in diastole phase and clinical outcomes	46
43	Correlations between first-order radiomics features of the left ventricle in diastole phase and clinical outcomes	47
44	Correlations between shape radiomics features of the left ventricle in diastole phase and clinical outcomes	48
45	Correlations between first-order radiomics features of the left ventricle in diastole phase and clinical outcomes	48
46	Correlations between GLCM texture radiomics features of the left ventricle in diastole phase and clinical outcomes	49
47	Correlations between shape radiomics features of the left ventricle in diastole phase and cardiovascular risk factors	50
48	Correlations between first-order radiomics features of the left ventricle in diastole phase and cardiovascular risk factors	51

1 Introduction

Nowadays, medicine is a field that generates incredibly large amounts of data coming from patient health analysis, investigations, electronic health records and many more. To give a tangible proof, 153 exabytes of medical data was generated worldwide in the year 2013. The amount of data increased gradually for the past years, and in the year 2020, it established a record of 2314 exabytes. The other problem besides the quantity is the distribution, because this data is distributed in a wide range of variables as well. All of these data can be processed and analysed for various purposes and one common goal: design health politics that reduces the probability of getting a disease.

The UK-BioBank is a large-scale biomedical database created and managed for research purposes. Due to being constantly growing and being accessible from everywhere around the globe with the pertinent authorization, it is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health. The database is currently focused on the research of common diseases that threaten most people's health. Even tho it has an enormous amount of data, the UK-BioBank doesn't dispose of tools to visualize it properly at the moment and investigators can't see intuitively the information this data bank has to offer.

With the right procedures, it's possible to go from raw information to applications that help medicine accomplish its objective. Today, some examples are real-time alerting, health planning, research of cures and vaccines for diseases and suicide and self-harm prevention. The reason behind this project is to create an application that gives a proper visualization tool to the UK-BioBank investigators.

With access to part of this data and some research, I developed a tool that will help investigators finding new relations and associations between physical characteristics and diseases through data visualization, focusing on heart structure features extracted from imaging called radiomics. Thanks to this associations, investigators can formulate new hypothesis on health politics and support them.

This data set is perfect for this work because it represents the average citizen with access to the health applications this tool could help developing, ensuring the utility of this instrument.

2 Objectives

The main objective of this work is developing a web page that uses the big data of the radiomics provided by the UK-BioBank to help investigators formulate hypothesis on the associations between human physical and mental traits and routine behaviours and the odds of getting a disease. The web page will also provide information about radiomics and the heart features they provide, which are key on this purpose due to the conclusions an investigator can draw on their relations with human health.

To accomplish this main goal, the following objectives are established:

1. Understanding the data set.
2. Research on data analysis methods for big data.
3. Design the tool for the data visualization using a simple web architecture.
4. Test the page with real users from the BCN-AIM lab.

These four objectives were created after a period of investigation and deliberation, ensuring the creation of a solid tool that will be used by investigators.

3 Planning

The planning for this project was set on October 2020. It was a general idea for the time distribution and wasn't an in depth schedule.

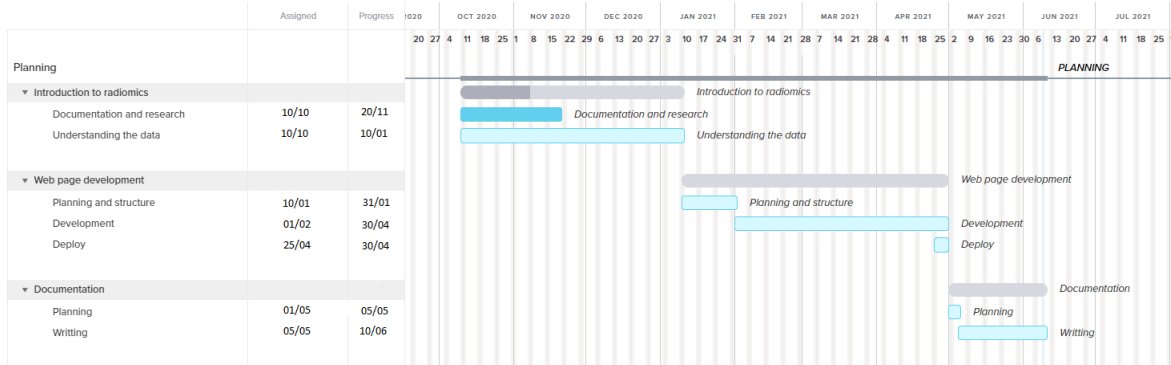


Figure 1: Initial planning

The first step of the planning is to learn about radiomics and the data involving the project. For this reason, a month and a half is assigned to research and documentation about radiomics and related concepts, and three months are set for understanding the data set and how it is extracted from the UK-BioBank.

Once the concepts needed for this project are acquired, the next step is to take some days in order to start thinking on the web page. Initially, the most important thing is deciding the software that will be used during the implementation. Next, structuring the web page and deciding the distribution of the information. With this in mind, the implementation can begin. This period is the longest on the planning and it has a duration of four months.

Finally, in the last days, the objective is the planning and writing of the documentation.

This planning hasn't been followed strictly and the schedule has changed, mainly in the web page development stage. It has taken an extra month to finish the RWAS web page due to some difficulties on the programming and deployment.

4 Theoretical framework

4.1 Radiomics

Radiomics is an image analysis technique that extracts quantitative features from voxel level data of medical images taken on routine-care checks. These features are stored in a minable structure with the objective of being used in the development of models that relate them to biological phenotypes. Thanks to the relations that radiomics bring to light, models for disease diagnosis and prediction become faster and more accurate.

Radiomics are extracted from designated regions of interest of cardio magnetic resonances of a large number of patients. The regions of interest are the parts of the image containing the most valuable information, such as the ventricles and the myocardium.

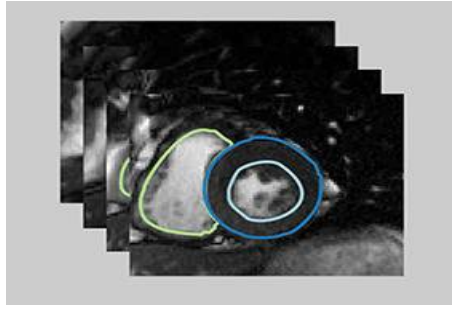


Figure 2: Example of a cardio magnetic resonance and the designed regions of interest

The features extracted from this technique can be divided in two categories: shape and signal-intensity.

Shape features, as its name indicates, are the ones based on the shape of the figure that has been rendered from the cardio magnetic resonance images. These characteristics are translated to geometrical quantifiers, such as volume and area of the surface, and descriptors of the overall shape, such as sphericity, elongation and compactness.

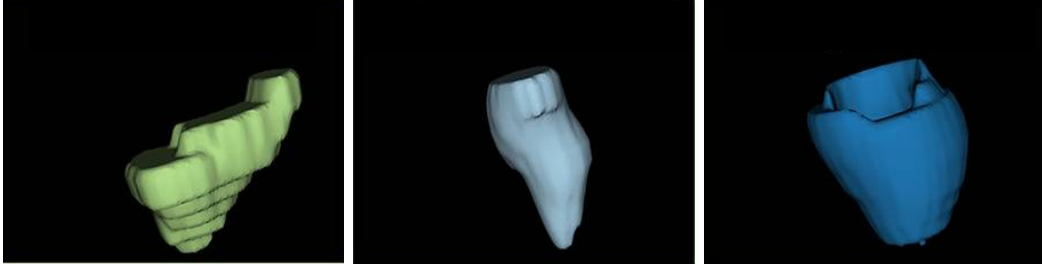


Figure 3: 3D renders of the left ventricle, right ventricle and left ventricle myocardium

Intensity-based radiomics can also be divided in two groups: features describing the global distribution and features describing patterns of voxel signal intensities.

The ones describing the global distributions are called first-order features, and they only consider the distribution of signals of each voxel individually. They are derived from histogram-based basic methods and display the intensity levels of each defined region of interest from every voxel as single quantifiers such as mean, median, maximum and minimum. There are also more quantifiers that describe more complex concepts, such as randomness, skewness and kurtosis. These three represent entropy, asymmetry and flatness respectively.

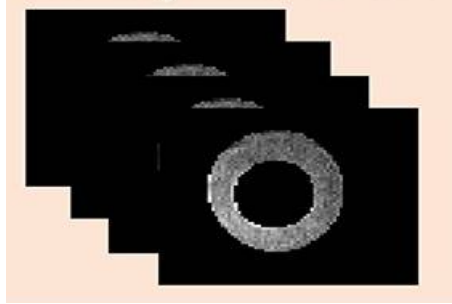


Figure 4: First-order radiomic of the left ventricle myocardium

The other type of signal-intensity features is texture features. This type of features are statistical descriptors of the relations between neighboring voxels of similar signal intensities. They are calculated using various matrix analysis methods according to standardized mathematical definitions. An example of matrix is the GLCM matrix, which describes the relationships between pairs of voxels within given distance and angle. Another example is the GLRLM matrix, which shows the number of consecutive voxels having the same intensity.

To summarize, radiomics is a technique that extracts a large amount of features from the regions of interest of cardio magnetic resonance images, which are divided in shape radiomics, first-order radiomics and texture radiomics. This information is later processed and analysed statistically to extract conclusions on the relations between the features and human traits and behaviours.

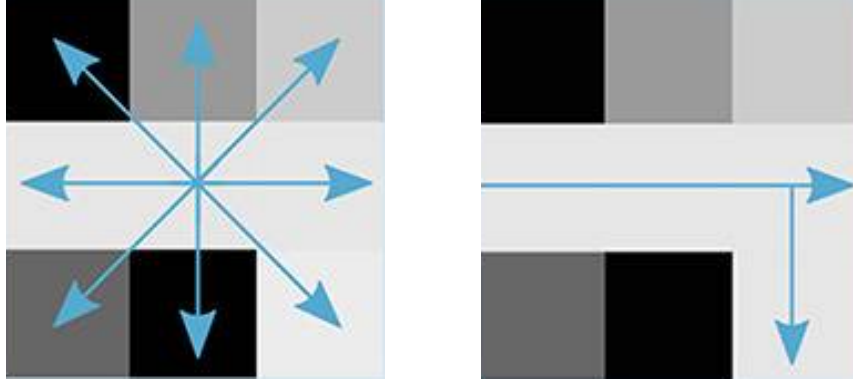


Figure 5: GLCM matrix and GLRLM matrix

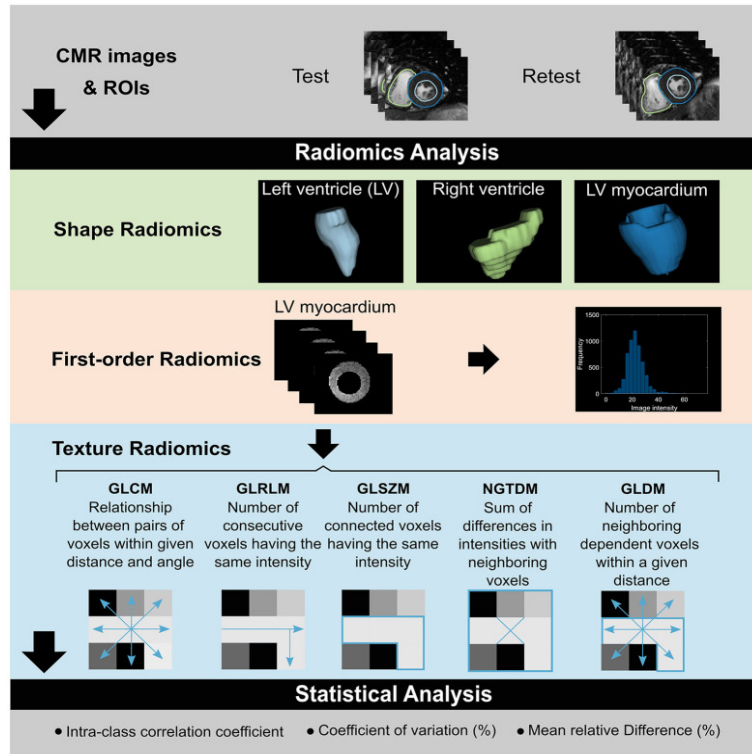


Figure 6: Graphical representation of the flux of radiomics processing extracted from *Repeatability of Cardiac Magnetic Resonance Radiomic*[1]

4.2 P-values

The p-value is a numerical value that reflects the degree of data compatibility with the null hypothesis in null hypothesis significant testing. If the p-value is lower than 0.05 it indicates a strong probability of the existence of the relation. From there, the lower the p-value, the most significant is the relation.

In radiomics, the amount of data extracted is really large. For this reason, finding out relations between the different features and diseases is a challenge. Formulating hypothesis and using the p-value threshold with the purpose of selecting

the meaningful results is the current standard and it has proven to work, but due to the dimensions of the data sets, the recommendation is to use a threshold of 5×10^8 instead of 0.05, avoiding a large amount of false-positives. This value hasn't been arbitrary selected as an artificial way of reducing the amount of positive results, but reflects the Bonferroni correction, a method used in statistics to reduce the total of positives when the number of hypotheses is large enough to provide too many of them.

In the case of the radiome wide association study web page, the p-value is used to represent only significant data on the plots so investigators can theorize about the relations between radiomic features and diseases without worrying about false-positives.

5 Methodology

5.1 Dataset: UK-BioBank

The UK-BioBank is the most detailed, long-term prospective health research study in the world, enabling the international scientific community to better understand a range of common and life-threatening diseases.

For my project, only access to a small part of the massive data is has collected over the last fifteen years was given. All of the handed data is contained in several static files in comma-separated values or excel table format. These files can be categorised by the following:

1. Diseases codes
2. Radiomics features for association analysis
3. Radiomics features for cardiovascular risk factors

The diseases codes files are raw data extracted from patients routine-care checks. These patients have agreed to allow the UK-BioBank to use their medical data for research purposes and it's confidential. The name of the documents is a number that represents the standard of the codes they contain, each one indicating a cardiovascular disorder.

The codes are set in ICD9, ICD10 and self-reported standards. ICD stands for International Classification of Diseases and refers to algorithmically defined outcome data. Self-reported refers to diseases reported by the patients on their on their routine-care checks, mainly in the first one. Each of these codes files contain a column with the code of the patient and a set of columns that indicates the point in time when the disease was diagnosed to the patient.

These files are examples of the type of files that contain the data for generating the radiomics features files.

The radiomics features files are a compilation of radiomics related values. In the case of the association analysis data, they are composed by odds ratios, corresponding p-values and confidence intervals for the odds ratios. These odds ratios describe the impact the constant effect of radiomics on the likelihood that an outcome will occur, and are calculated with a logistic regression model. The files for calculating the cardiovascular risks factors contain the beta-coefficients, p-values and confidence intervals that describe the associations between these risk factors and each radiomic feature.

5.2 Technologies used

The design of the radiome wide association study web page, or RWAS web page to abbreviate, has the objective of begin a site up to nowadays standards. Like most of the data visualization tool, the main focus of the Radiome wide association study web page is plotting the information in an understandable and intuitive way. To achieve this objective, it uses a selection of technologies that allows the web page to be a tool a researcher would like to use to develop his theories.

HTML

Today, every web page uses HTML. The HTML acronym stands for HyperText Markup Language, and it's a technology designed to display documents in a web browser. This programming language is composed by a series of tags that define the elements of a web page and their distribution along the document. For example, it defines text, images, fonts, sizes of every element, lists and colours.

The web page it's conformed by a total of six HTML documents. Each one of these documents has a different objective and represent a part of the web site. Specifically, one of them displays the home page, four of them are for the different type of plots and the last one stands for the contact page.

These files also contain a series of technologies besides HTML in charge of defining the styling, the selection of the data the user wants to be displayed and the plotting of this same data.

CSS

CSS or Cascading Style Sheets is a programming language for styling the presentation of HTML documents. It has the power of making a web page look attractive and intuitive for the user, describing the structure and distribution of the elements defined in the HTML document and giving them distinctive good looking styles.

Every single one of the HTML files of the RWAS web page web page has CSS styling. Thanks to this technology, the sections of the site share a common dark and professional theme, giving a sensation of cohesion and unity to the user.

JavaScript

The JavaScript technology is a high-level object-oriented programming language compatible with all web browsers, and has the objective of adding interactivity to web pages. For example, it can be used to add events to buttons, refresh only certain parts of the page or display alerts for the user.

Thanks to this powerful technology, a series of scripts in the HTML documents

manage the interactivity between the user and the RWAS web page, giving the sensation of an interactive platform and not only a group of plain text and images. In combination with other technologies, JavaScript makes the navigation easy, improves the user experience and displays the different types of plots the site has to offer.

Flask

Flask is a python framework that communicates perfectly the back-end and front-end applications. It's also denominated a microframework for its simplicity and small core, but it has the potential to expand through adding dependencies. Is the framework the RWAS web page is running on, and it has allowed me as a developer to have only the tools the site needs, without extra not needed features.

JQuery

There are a lot of libraries for JavaScript, and JQuery is one the most popular. Its main objective is to make the programming easier for the developer. It helps with the manipulation of the document object model and the event handling.

Google charts

Google charts is another JavaScript library designed specifically for data plotting with a wide range of chart types. Some examples are bars, histograms, bubbles, columns, lines and areas. All of them are based on HTML5 technology, simple to use and easy to understand.

This is the library in charge of displaying the histogram of the data showcase section in the RWAS web page.

Plotly

Plotly is also a plot-oriented library, with a section dedicated only to JavaScript. It has an even larger range of interactive plot types than Google charts, and it also has more customization options.

This technology is used in the odds section of the RWAS site and displays a Manhattan plot, a modification of a scatter plot that manages data with a large number of data-points with a distribution of higher-magnitude values. It is also present in the correlations sections displaying a heatmap.

Holoviews

The Holoviews library is a plotting library less powerful than the ones mentioned before, but it has a chords plot. This plot is the one in charge of displaying the relations in the cardiovascular risk factors section of the RWAS web page.

5.3 Radiome wide association study web page

5.3.1 Homepage

The homepage is the very first thing any user will see upon accessing the site. It serves as an introduction to radiomics and its capabilities, and it's also a portal to all the sections of the web page.

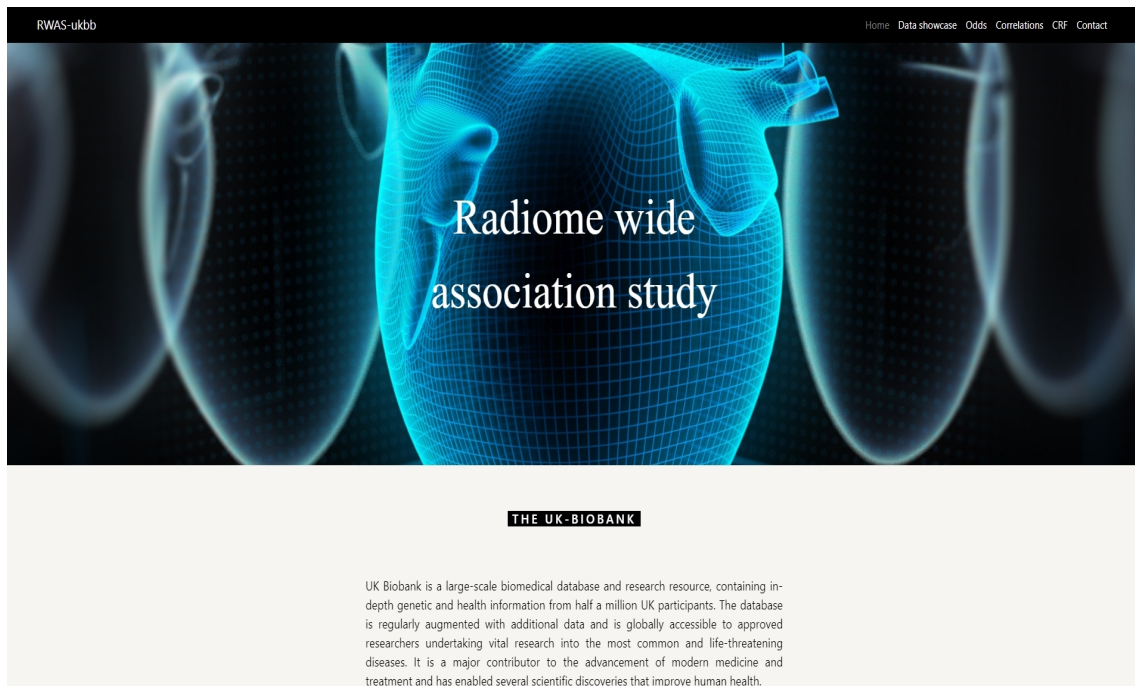


Figure 7: Homepage of radiome wide association study web page

On the top side there is the navigation bar that has to be used to navigate through the page. It has six different sections: Home, Data showcase, Odds, Correlations, CRF and Contact. It indicates the current section the user is in by shadowing it. For example, in the figure above Home is shadowed, as it is the current page. On the left side there is the abbreviation of the name of the web page, RWAS-ukbb, referencing Radiome wide association study - UK Biobank. This can be used to navigate to the homepage as well.

The center of the page displays its name, and below it there is a brief explanation for the UK-BioBank, radiomics, and the main goal of the web page. With this homepage, the user can learn in a few words the basics of the page and the concepts it works with.

5.3.2 Associations with developmental, lifestyle, clinical and environmental influences

The objective of this section of the web page is to display information about the relation of certain human characteristics, traits behaviours and the odds of getting a disease.

In this case, the site has two separated sections for plotting the two types of data sets related to this concept: Data showcase and Odds. Each one has a clear plot that suits the data, making easy for the user to understand it.

Data showcase

This part of the RWAS site has a side bar where the user inputs the information he requires and a histogram plot displaying it.

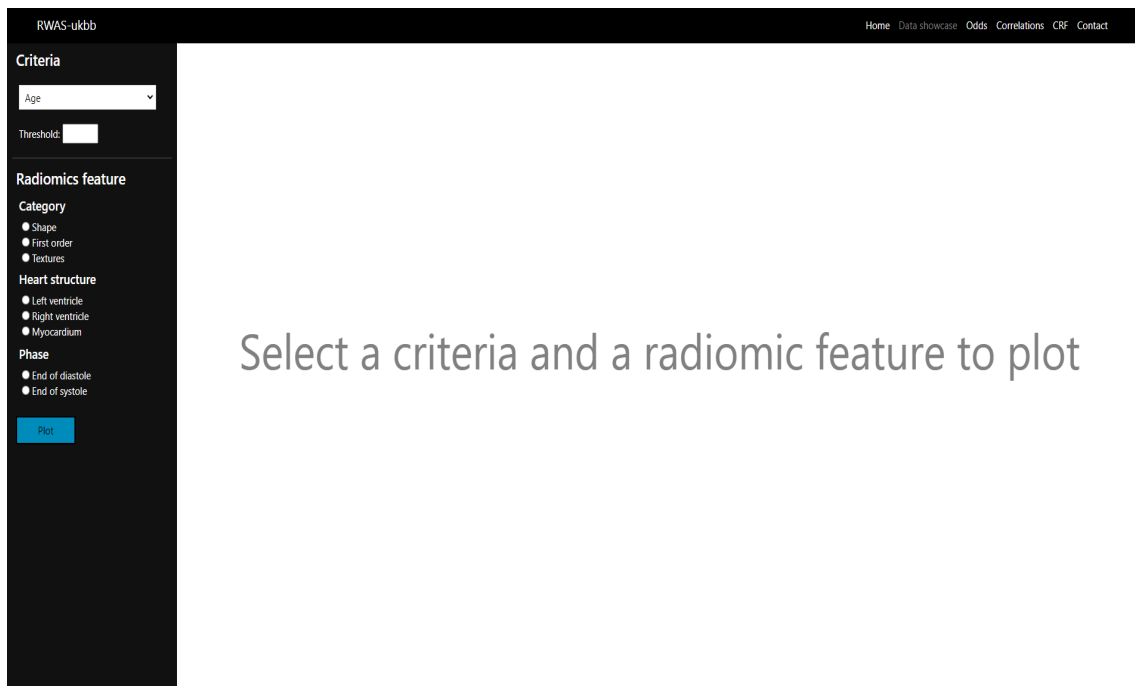


Figure 8: Data showcase section without selection

On this figure the side bar draws the attention immediately. It has two groups of inputs: criteria and radiomics feature. The criteria is the characteristic the user wants to see the radiomics feature on, and the radiomics feature itself is the field which values are represented in the histogram. At the bottom, there is a button for plotting the histogram once the inputs are introduced.

The criteria input can be binary or numeric. When a numeric feature is selected, an additional input field appears telling the user to specify the threshold of the fea-

ture he wants to work with. If the selected feature is binary, the threshold input isn't shown. The radiomics feature needs from four to five inputs to be determined:

1. **Category:** a radiomics feature can be divided in shape, first-order or texture.
2. **Category subfield:** on selecting a category, a dropdown like the one on the criteria input will appear. It contains all the radiomics features the selected category has.
3. **Texture subtype:** if the textures category is selected, a group of five sub-categories will pop up: glcm, glszm, glrlm, ngtdm and gldm. Only when one of these is selected that the dropdown with the categories will appear.
4. **Heart structure:** the radiomics category is focused in one of these three parts of the heart structure: the left ventricle, the right ventricle and the myocardium.
5. **Phase:** the radiomics features are extracted in a specific moment of the heart phase. It can be when the heart has finished the diastole or when it has finished the systole.

Once introduced the inputs and clicked the Plot! button, the histogram will appear.

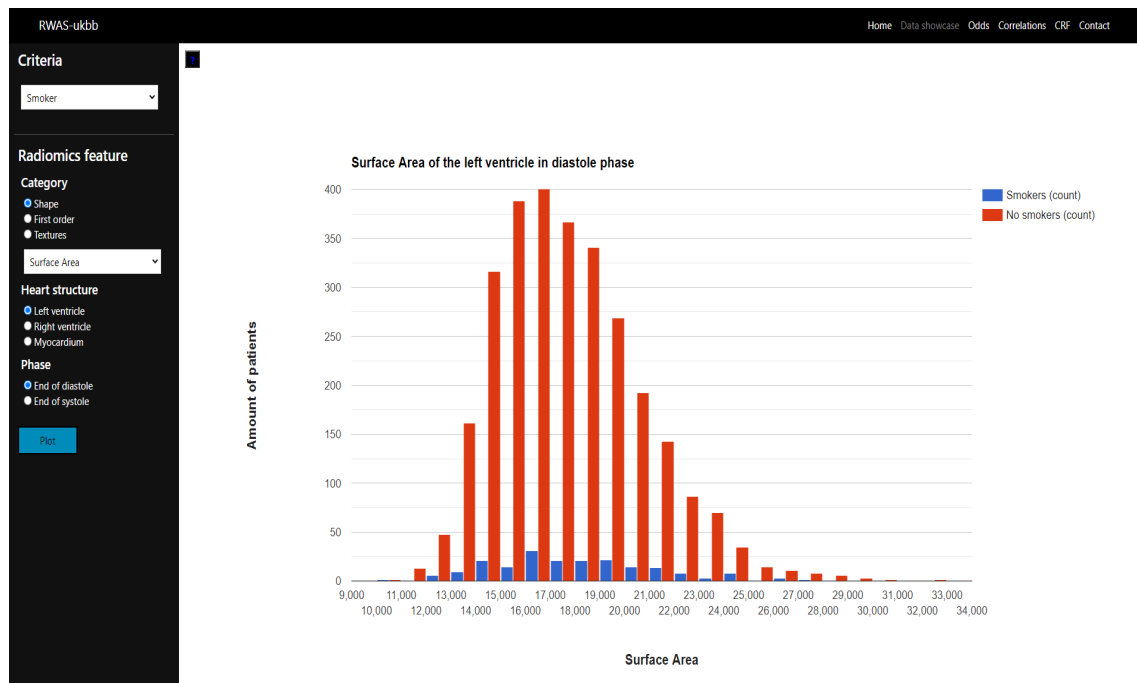


Figure 9: Data showcase section with the histogram

In the case of the figure above, the user wanted to know the surface area of the left ventricle in the end of diastole for smokers and non smokers. The histogram

shows perfectly how the surface area is much bigger for non smoking patients, which means that, as we all know, smoking affects severely on the state of the heart. The y axis always corresponds to the count of patients on each interval, and the x axis corresponds to the range values the radiomics feature can be in.

Finally, at the top-left corner of the histogram there is a help button. On clicking it, the user will see a model pop up and show a brief explanation of the plot and how it works.

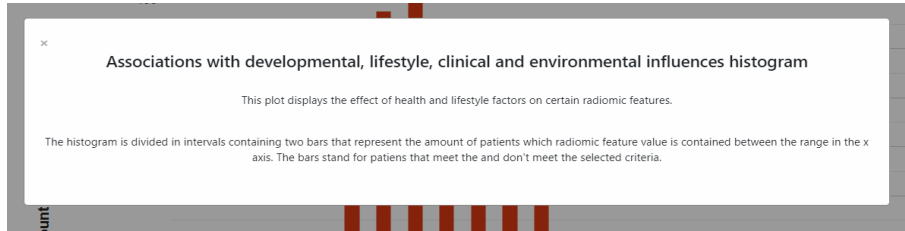


Figure 10: Help pop up of the Data showcase section

Odds

The general look of the Odds section is very similar to the Data showcase section. The only difference is the side bar, which has different parameters and inputs.

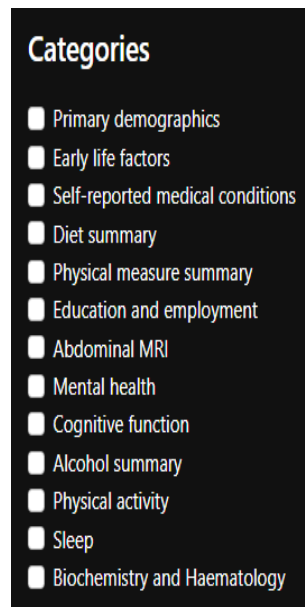


Figure 11: Side bar for the Odds section's field selection

Each one of the options is an group of human traits and behaviours with common characteristics. They have the following meanings:

1. **Primary demographics:** attributes related to the population the patient is part of. This is interesting to detect correlations between the characteristics of certain populations and the odds of getting a disease.
2. **Early life factors:** life conditions a patient has suffered in his early life, mainly before the eighteen years old boundary. Some examples are chronic diseases, hospitalization, living on a village and parental unconcern.
3. **Self-reported medical conditions:** medical conditions the patient has reported by itself.
4. **Diet summary:** attributes of the usual diet of the patient, such as calorie consumption and the variability of the food.
5. **Physical measure summary:** quantitative information on physical conditions and properties of the patient for essential activities.
6. **Education and employment:** historical of the patient's education and employment career. This is meant to determine a relation between behaviours that help getting specific diseases and the educational and professional career.
7. **Abdominal MRI:** features extracted from the imaging of abdominal magnetic resonances.
8. **Mental health:** past and current mental health conditions of the patient, such as cognitive disorders and neurodegenerative disorders.
9. **Cognitive functions:** cognitive capabilities of the patient.
10. **Alcohol summary:** past and current behaviors related to alcohol consume.
11. **Physical activity:** usual physical activity of the patient.
12. **Sleep:** sleep characteristics of the patient, such as time and consistency.
13. **Biochemistry and Haematology:** biochemistry refers to the analysis of the chemistry of the compounds in the human body, catabolism and metabolism and haematology refers to the attributes of the blood and the blood-producing organs.

Once the user has selected the fields he needs, a manhattan plot is displayed, showing the relations between diverse diseases and the fields.

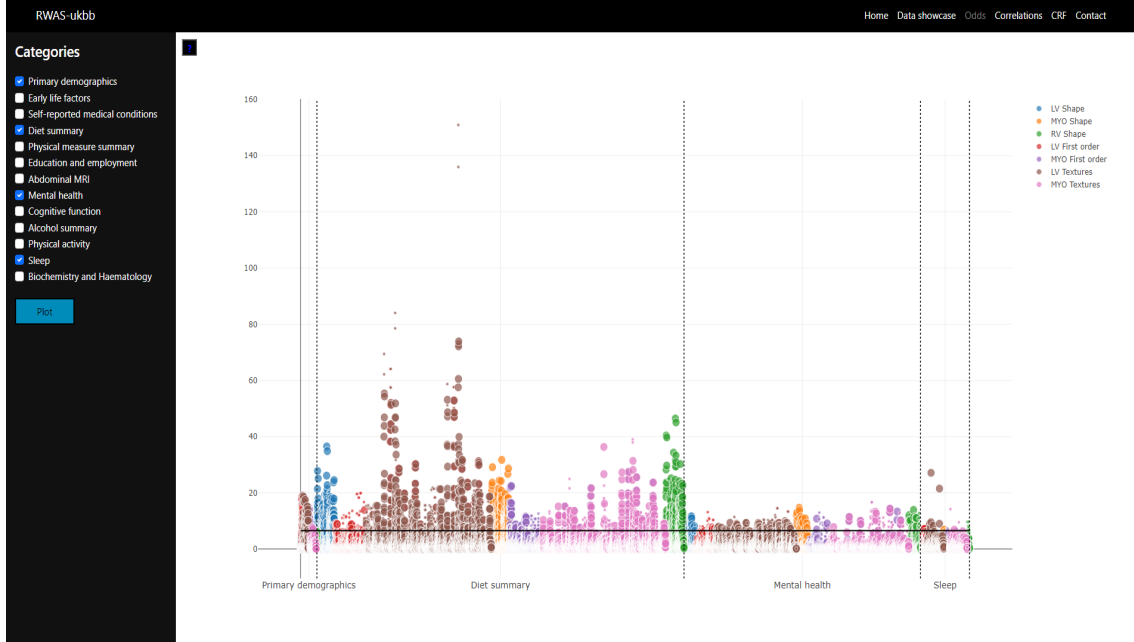


Figure 12: Manhattan plot in the Odds section with selected fields

This graph is composed by two axis: the y axis indicating the relevance of the dot and the x axis indicating the group which the dot belongs to. The value represented in the y axis is the negative logarithm to base 10 of the p-value calculated for the individual disease. The reason behind this is to relate the height of the plotted dot and it's relevance by its assigned p-value. There are also three sizes for the dots relative to their correlation value. The black horizontal line indicates the Bonferroni value, making a distinction between the features that are meaningful and the one that aren't.

The Plotly library's graphs are interactive and allow the user to navigate thorough them, zoom in and out to focus on certain parts, toggle only the dots that belong to specific groups in the legend and hover over them for more information. This manhattan plot is really dependant on this features due to the amount of data that displays in the space it disposes of. A big part of the data appears only when the mouse hovers a dot, and it's relative only to that dot.

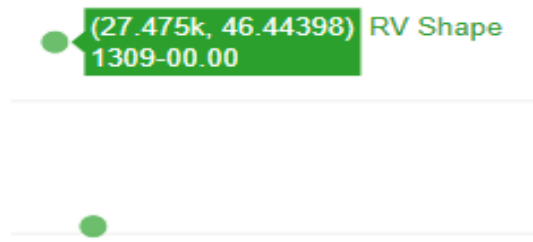


Figure 13: Information displayed on hoverign a dot in the manhattan plot

The first value on the parenthesis indicates the x value, which only purpose is to group the plots in the same category. The second value indicates the y value, which is the resultant value of the negative logarithm to base 10 of the p-value the disease has assigned. In other words, it indicates the importance of the effect of the disease on the human traits and behaviours that are grouped in the category the dot belongs to. The colour and the text on the right indicate the category the radiomics feature pertain, and the number in the bottom part is the specific disease code. In the case of the figure above, the dot represents a disease related with a radiomics feature extracted from the right ventricle, it's related to the shape and has the code 1309-00.00.

As in every section of the site, this section has a help button on the top-left side where the user can see a brief explanation of how does the manhattan plot work and how to interact with it.

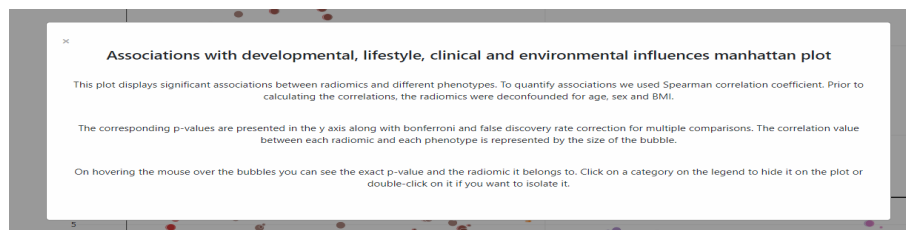


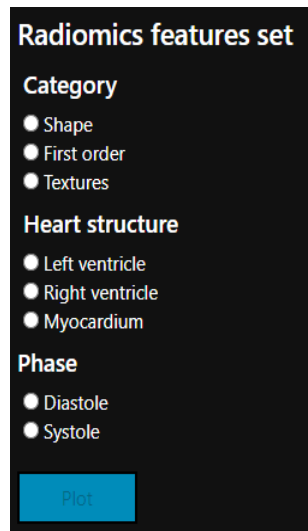
Figure 14: Help pop up of the Odds section

5.3.3 Associations of radiomics with clinical outcomes

The Correlations section of the RWAS web page displays the odds ratio of the radiomics features on clinical outcomes. In this particular case, the odd ratio represents the constant effect of radiomics on the likelihood of the occurrence of a clinical outcome. The chosen plot type for this case is the heatmap, a graphical representation that uses a colour system to give value to its elements.

The clinical outcomes used in this section have been selected for the high impact they have on the population of the United Kingdom, which represents every population with similar life level and conditions. The selected outcomes are asthma,

bronchitis, cardiac diseases, chronic obstructive pulmonary disease, depression, diabetes, high cholesterol, hypertension, peripheral vascular disease and all causes mortality or demise. In this section, the side bar is simpler and has less options.



Radiomics features set

Category

- ☐ Shape
- ☐ First order
- ☐ Textures

Heart structure

- ☐ Left ventricle
- ☐ Right ventricle
- ☐ Myocardium

Phase

- ☐ Diastole
- ☐ Systole

Plot

Figure 15: Side bar for the Correlations section's field selection

Similarly to the Data showcase section, the user has to select a set of radiomics features grouped by their category, heart structure and phase. Once selected, the heatmap will be plotted with the associations between each one of the radiomics features of the selected group and the clinical outcomes.

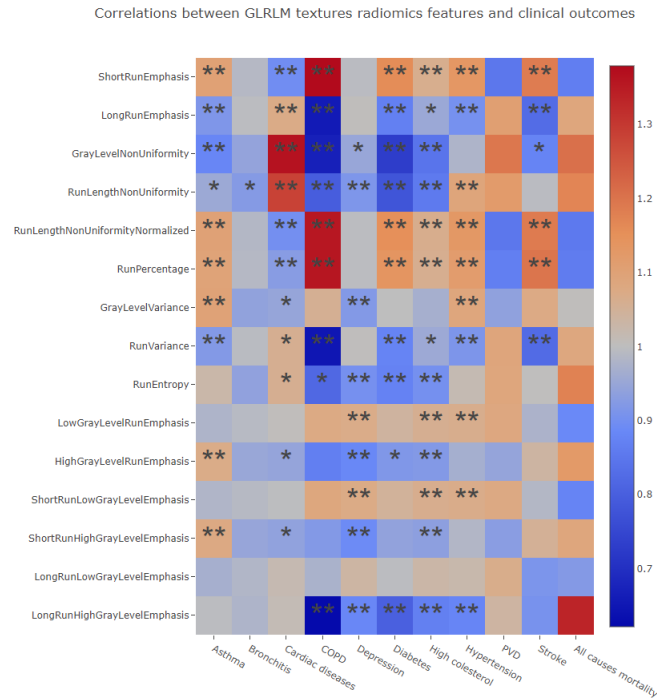


Figure 16: Heatmap in the Correlations section with selected fields

In this heatmap, the rows represent the radiomics features and the columns the clinical outcomes. As the bar on the left side indicates, the range of values the correlation can variate on is from 0.5 to 1.5, represented by a gradient of colours. This gradient has a dark blue colour as the lowest possible value, and red as the highest.

Each square shows how much a radiomics feature and a clinical outcome are associated through the color, and the reliability of the results through the number of asterisks. A square can have none, one or two asterisks, being none the least reliable and two the most reliable. This reliability is given by the p-value the association has assigned, which has been previously calculated.

The help pop up of the Correlations section explains the data displayed and how to understand the heatmap.

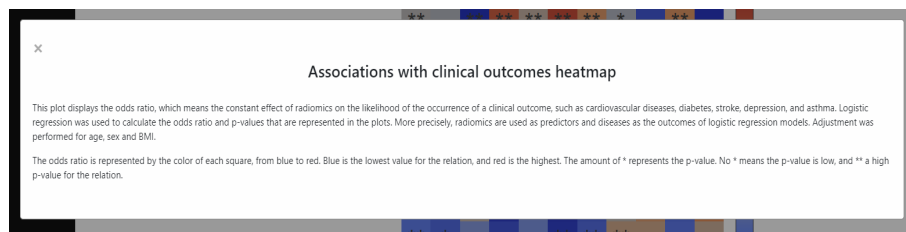


Figure 17: Help pop up of the Correlations section

5.3.4 Association of radiomics with cardiovascular risk factors

A cardiovascular risk factor is a measurable characteristic that is causally associated with increased heart diseases frequency, and also acts as a predictor for a higher chance of getting said diseases. In this section of the RWAS web page, the user can look for associations between these risk factors and each radiomics feature.

This time, the side bar has a group of cardiovascular risk factors besides the usual radiomics features group selection fields.

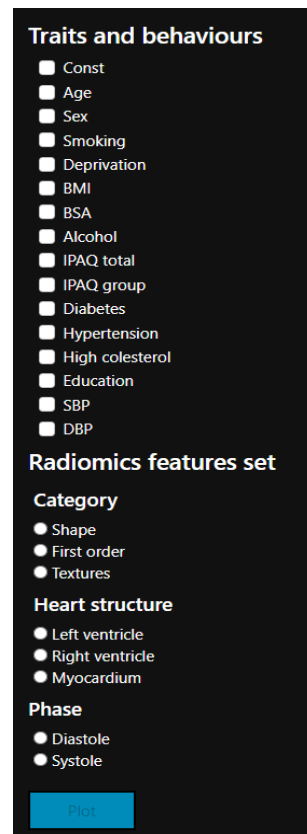


Figure 18: Side bar for the Cardiovascular risk factors section's field selection

The traits and behaviours available have been selected by the same reasons the clinical outcomes were in the Correlations section. Each one of them has a strong impact in the mean population due to being common and significantly increasing the possibility of getting a heart disease. The selected cardiovascular risk factors are the following:

1. **Age:** the age of the patient has a high impact in the odds of getting a heart disease because the heart weakens with the years.
2. **Sex:** it is proved that men are more likely to develop a heart disease at an earlier age than women.

3. **Smoking:** most of the substances in tobacco are harmful for the blood vessels.
4. **Deprivation:** socioeconomic deprivation refers to the difference between individuals in accessing economic, social or material resources.
5. **BMI:** the BMI acronym stands for body mass index and measures the body fat in relation to the height and weight. It is well known that an excessive amount of fat greatly increases the odds of heart disease, such as diabetes and high blood pressure.
6. **BSA:** the BSA acronym stands for body surface area and it's a numerical descriptor of the body. Therefore, is also a predictor for heart diseases.
7. **Alcohol:** past and current behaviors related to alcohol consume. A high alcohol consume increases the cholesterol and blood pressure levels, and contributes to weight gain.
8. **IPAQ:** the IPAQ or International Physical Activity Questionnaire is a questionnaire developed to measure health-related physical activity in populations. Regular physical activity reduces high blood pressure, cholesterol levels and weight.
9. **Diabetes:** diabetes is a lifelong condition that causes high blood sugar level and damage blood vessels. It is also associated with obesity.
10. **Hypertension:** high blood pressure, also known as hypertension, is one of the most important cardiovascular risk factors due to the high deterioration it cases on blood vessels.
11. **Education:** several studies have demonstrated that a higher education level reduces the odds of getting a heart disease because of the behaviors it promotes.
12. **SBP:** systolic blood pressure is a measure that indicates the force at which the heart pumps blood around the body. High SBP is an indicator for hypertension.
13. **DBP:** diastolic blood pressure measures the pressure on the walls of the arteries between heartbeats. It is an indicator for hypertension as well.

Once the user selects the cardiovascular risk factors and the radiomics features group, a chords plot is displayed.

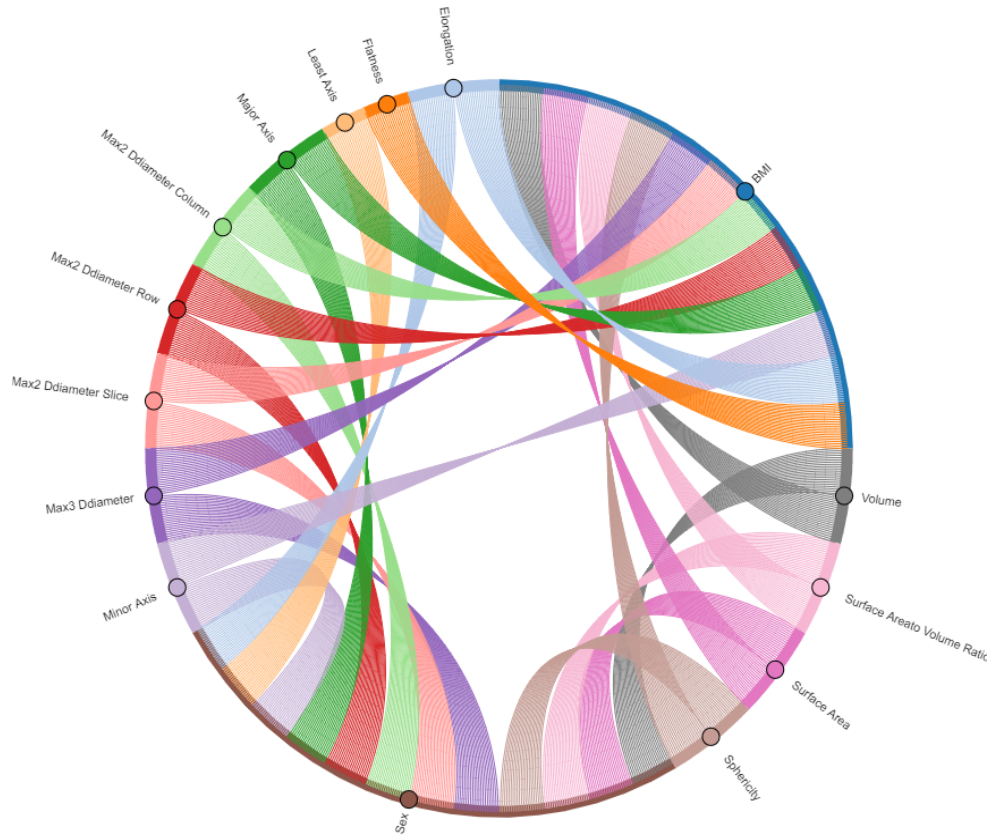


Figure 19: Chords plot of the Cardiovascular risk factors section for sex, smoking, BMI and shape radiomics features of the right ventricle in diastole phase

As the figure above shows, a chords plot is a graphical representation for visualizing data that describe relationships. The graph is composed of a circular figure divided in sections, each of which represents a field. From each field, a set of chords heads to other fields the origin field is related to.

Thanks to this section, investigators can visualize in a very intuitive way the associations between the risk factors he desires and a set of radiomics features. Because it is so easy to understand, the help pop up for this graph is very concise.

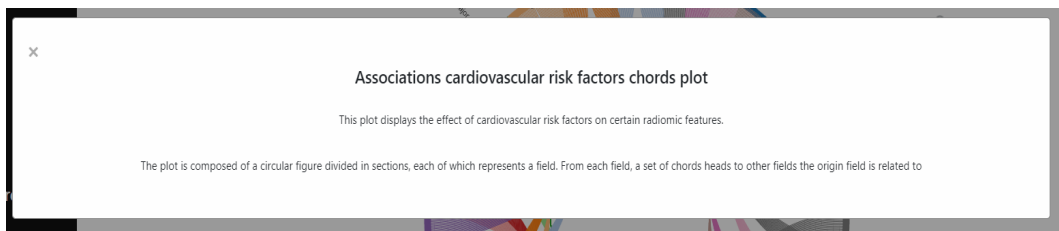


Figure 20: Help pop up of the Cardiovascular risk factors section

5.4 Deployment

In the world of web development, deployment means migrating an application to a web server, making it accessible by anyone from anywhere. It is usually the last step in the creation of a web page, and it takes place when the application is tested and ready to use.

Every deployment needs a server, a physical place where the resources are stored and accessible. For this project, the web site is hosted by Heroku. Heroku is a cloud application platform used world wide for multiple types of application profiles, from small amateur projects to big professional sites with a great amount of traffic. These profiles are also called dynos, and for presentation purposes, the Radiome wide association web page uses the free Heroku dyno because it is enough to show the potential of the tool.

A dyno is a lightweight container, an isolated Linux environment that provides computation, memory, an operative system, and an ephemeral file system for storing data. This container is a great system for protecting the web site, and also makes it scaling and flexible. Heroku's free dyno has 512MB of RAM and allows deployment directly from Git Hub, which is perfect.

6 Use cases

There are six general actions an investigator can perform in the RWAS web page.

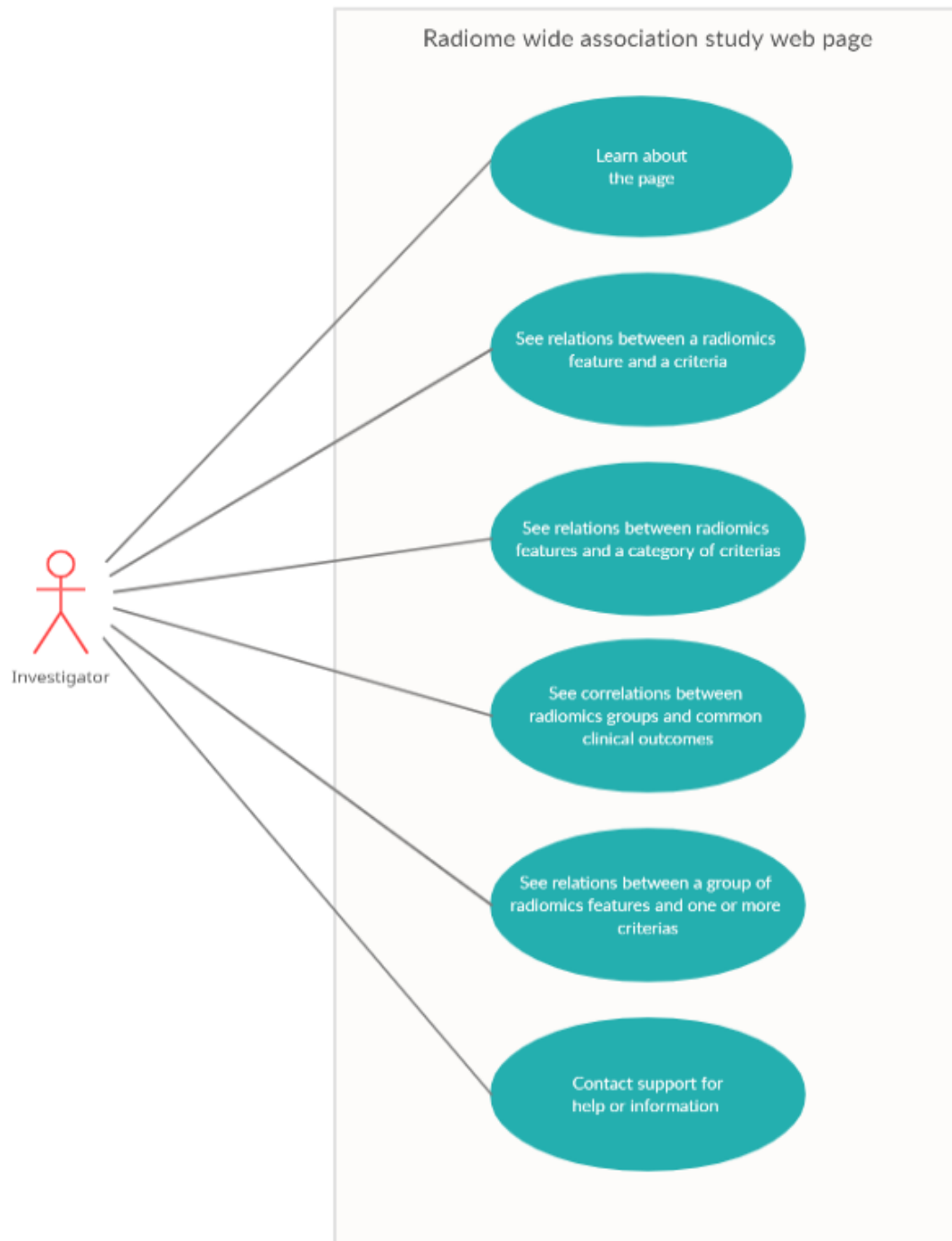


Figure 21: Use cases for the Radiome wide association study web page

Each one of these use cases corresponds to a section of the web page because

an action itself is simple and similar for every section. For example, using the Correlations section to display the correlations between shape radiomics features for the left ventricle in diastole phase or between first-order radiomics features for the right ventricle in systole phase and the clinical outcomes requires the same procedure. The steps for completing every use case are the following:

Use case 1: Learn about the page
Objective: Gather information about the required concepts to understand the plots.
Actor: Investigator
Procedure: 1. If not in the home page, use the navigation bar and click on "Home" or "RWAS-ukbb". 2. Scroll down the home page and read the information about the UK-BioBank, radiomics and the goal of the RWAS web site.

Use case 2: See relations between a radiomics feature and a criteria.
Objective: See a graphical representation of the relations between the radiomics feature and the criteria I choose.
Actor: Investigator
Procedure: 1. If not in the Data showcase section, use the navigation bar and click on "Data showcase". 2. Select a criteria using the Criteria dropdown on the left side bar. 3. Select a radiomics feature category using the Category radio selectors. 4. Select a heart structure using the Heart structure radio selectors. 5. Select a heart phase using the Phase radio selectors. 6. Click on the "Plot!" button.

Use case 3: See relations between radiomics features and a category of criterias.
Objective: See a graphical representation of the relations between the radiomics features and the categories I choose.
Actor: Investigator
Procedure: 1. If not in the Odds section, use the navigation bar and click on "Odds". 2. Select the categories using the Category checkboxes on the left side bar. 3. Click on the "Plot!" button.

Use case 4: See correlations between radiomics groups and common clinical outcomes.
Objective: See a graphical representation of the correlations between the group of radiomics features i choose and common clinical outcomes.
Actor: Investigator
Procedure: <ol style="list-style-type: none"> 1. If not in the Correlations section, use the navigation bar and click on "Correlations". 2. Select a radiomics feature category using the Category radio selectors. 3. Select a heart structure using the Heart structure radio selectors. 4. Select a heart phase using the Phase radio selectors. 5. Click on the "Plot!" button.

Use case 5: See relations between one or more criterias and a group of radiomics features.
Objective: See a graphical representation of the relations between the criterias: and the group of radiomics features I choose.
Actor: Investigator
Procedure: <ol style="list-style-type: none"> 1. If not in the Cardiovascular risk factors section, use the navigation bar and click on "CRF". 2. Select one or more criterias using the Criteria checkboxes on the left side bar. 3. Select a radiomics feature category using the Category radio selectors. 4. Select a heart structure using the Heart structure radio selectors. 5. Select a heart phase using the Phase radio selectors. 6. Click on the "Plot!" button.

Use case 6: Contact support to get help or information.
Objective: Obtain a contact mail adress or link.
Actor: Investigator
Procedure: <ol style="list-style-type: none"> 1. Use the navigation bar and click on "Contact". 2. Use one of the mail addresses or web pages to get help or information.

7 Examples of use

The Radiome wide association web page is a tool medical investigators can use to design new health politics that will lower the risks of getting certain diseases. To do so, the web page allows them to visualize relations between radiomics features, habits, physical and mental characteristics and diseases. Here there are some examples of the use an investigator could give to the RWAS web page.

Do males and females need different health politics to prevent heart diseases?

I want to design health politics to prevent heart diseases, such as strokes or arrhythmia. First of all, I want to know which features of the heart are correlated with heart diseases. With this objective in mind, I navigate to the Correlations section and display the correlations between different radiomics features and clinical outcomes.

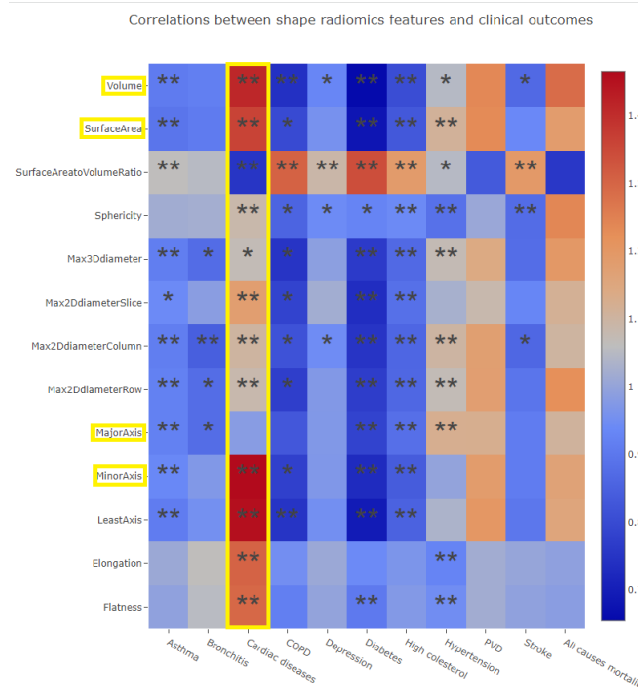


Figure 22: Correlations between shape radiomics features of the left ventricle in end of diastole phase and clinical outcomes

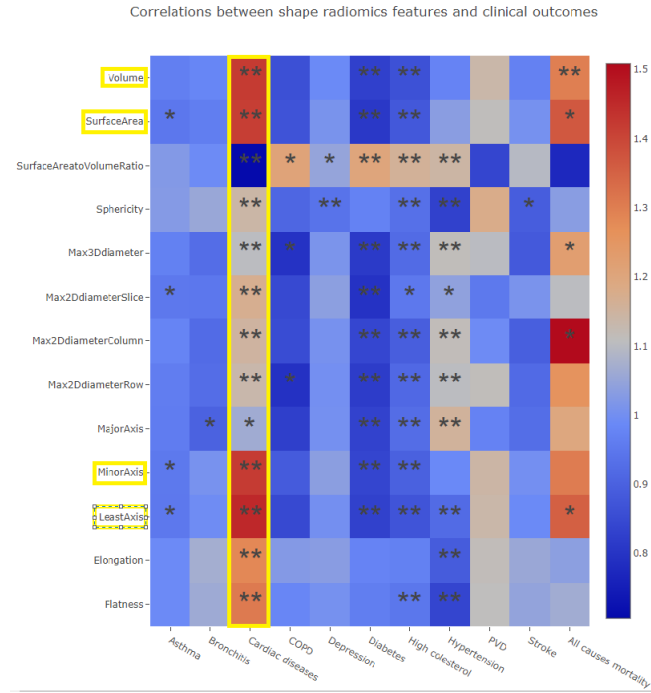


Figure 23: Correlations between shape radiomics features of the left ventricle in end of systole phase and clinical outcomes

With these two heatmaps, I can see there is a strong correlation between the volume, surface area, major axis and least axis of the left ventricle in both heath phases. Next, I want to know if this correlation exists for the right ventricle and myocardium as well.

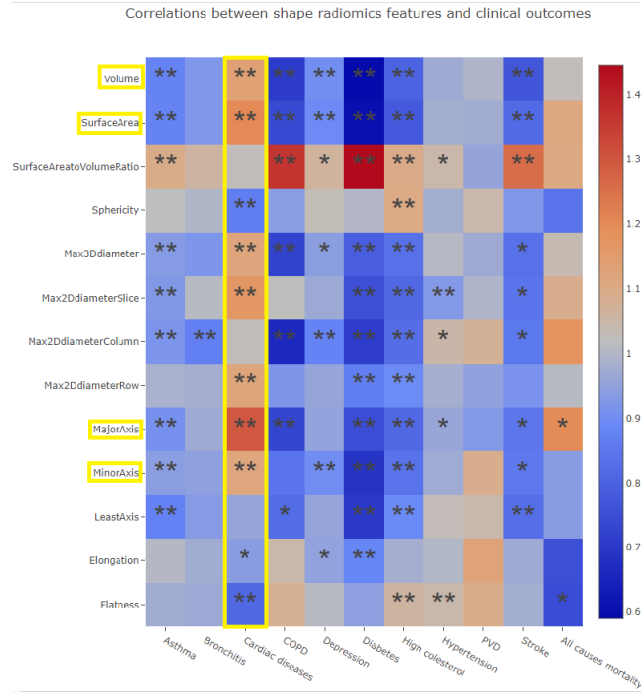


Figure 24: Correlations between shape radiomics features of the right ventricle in end of diastole phase and clinical outcomes

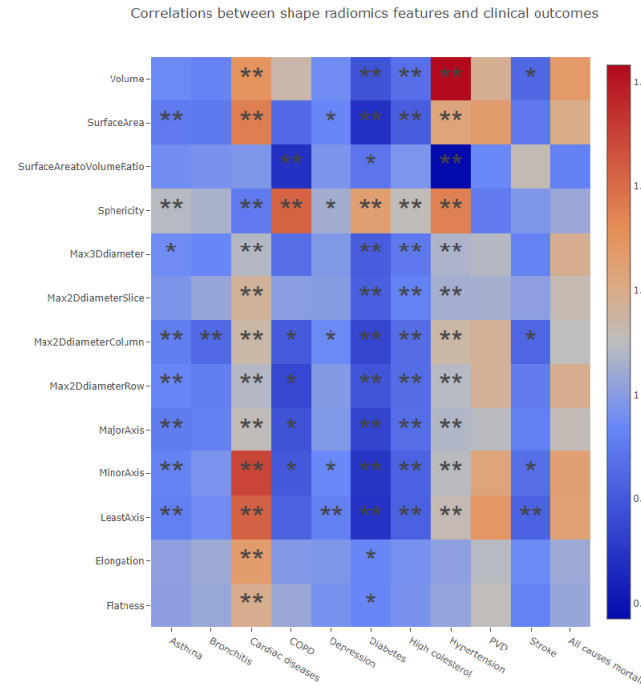


Figure 25: Correlations between shape radiomics features of the myocardium in end of diastole phase and clinical outcomes

With the two heatmaps above I can confirm that all of the shape of the heart is correlated with the odds of getting a heart disease. Knowing this and as the final step, I need to know if the shape of the heart, mainly the volume, surface

area, major axis and least axis, are different between males and females. To get the information required, I navigate to the Data showcase section, where I can see differences in radiomics features based on a criteria. In my case, the criteria I need to use is the sex of the patients.

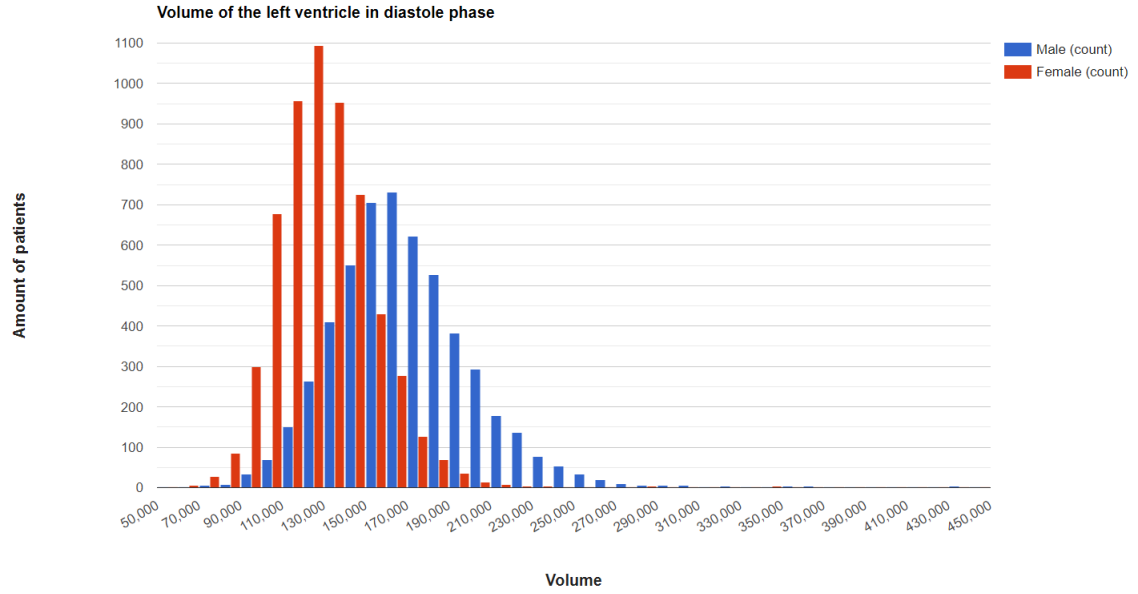


Figure 26: Volume of the left ventricle in diastole phase for males and females

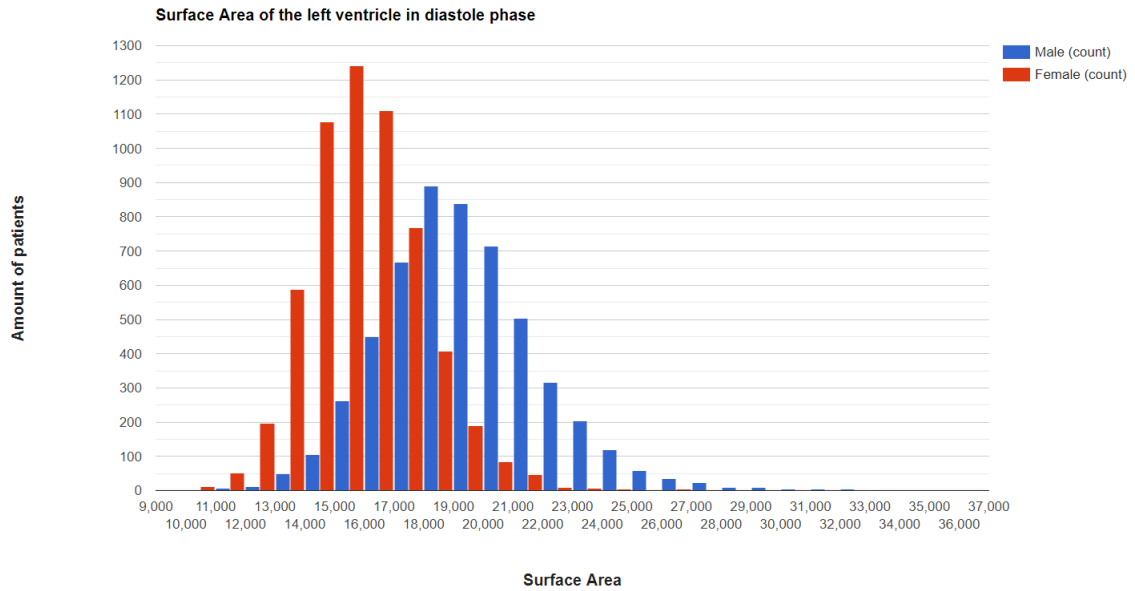


Figure 27: Surface area of the left ventricle in diastole phase for males and females

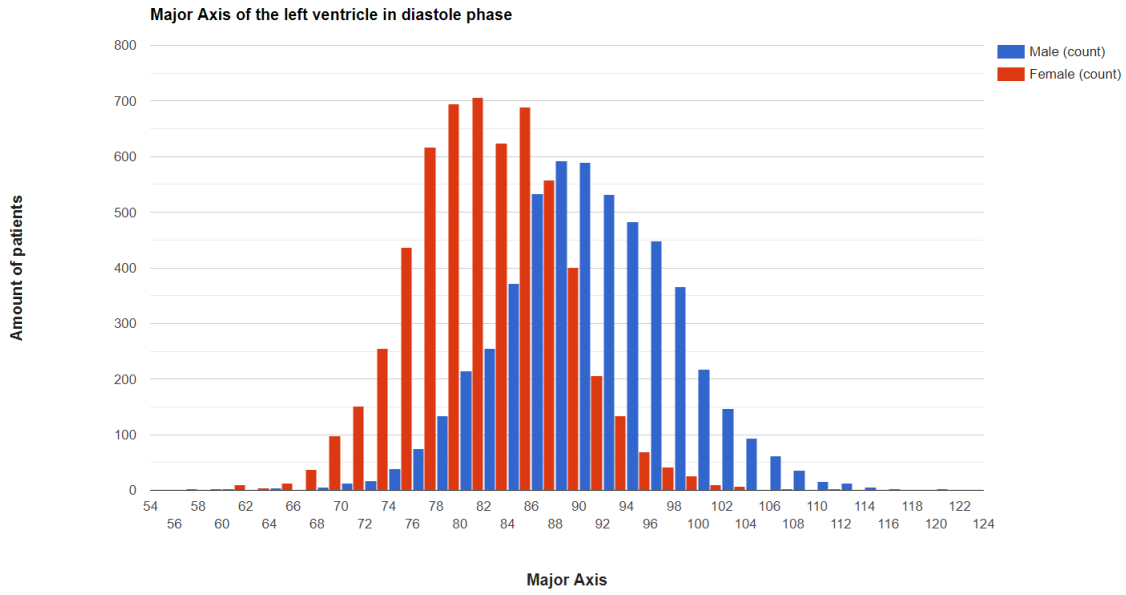


Figure 28: Major axis of the left ventricle in diastole phase for males and females

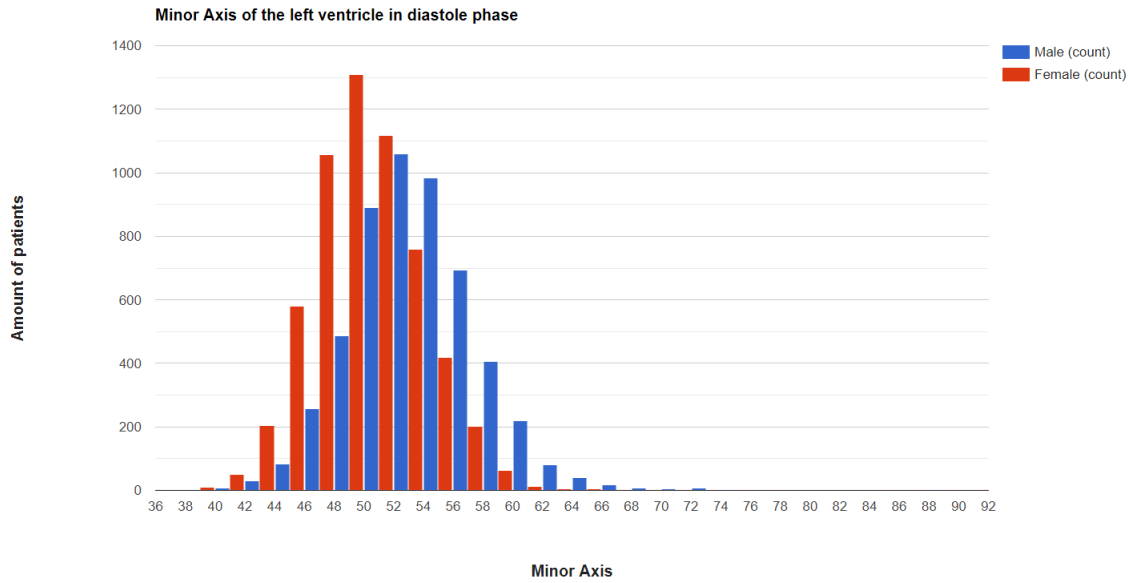


Figure 29: Least axis of the left ventricle in diastole phase for males and females

Thanks to this figures I can conclude that the shape of the heart is different between males and females. Generally speaking, man have a bigger heart than woman, and these shape features are directly correlated with the odds of getting a heart disease. This means I need to design different health politics for man and woman to decrease the odds of getting a heart disease.

As described in the article *Sex differences in first-ever acute stroke*[4], sex is a factor in the odds of suffering a stroke. Woman have a higher probability, but the mortality rate is higher for man.

Does smoking increase the odds of getting asthma?

I want to know if the habit of smoking is a determinant factor in the development of asthma in adults. Asthma is mostly known as a disease that develops in younger ages, but there are cases of adults developing asthma. First of all, I need to know which radiomics features are correlated with this disease.

After looking for correlations, I find that most of the right ventricle radiomics features corresponding to the first-order category have a high correlation value, and also a high p-value. This is true for both end of diastole and end of systole phases.

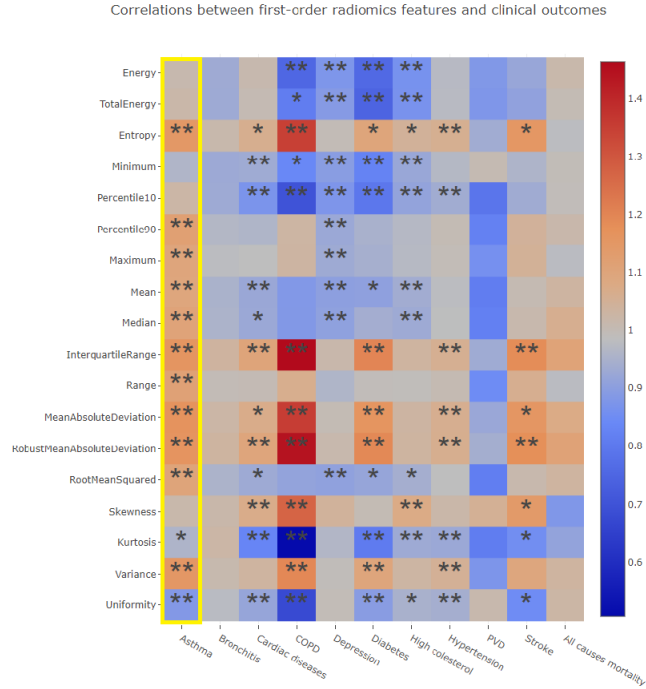


Figure 30: Correlations between first-order radiomics features of the right ventricle in the end of diastole phase and clinical outcomes

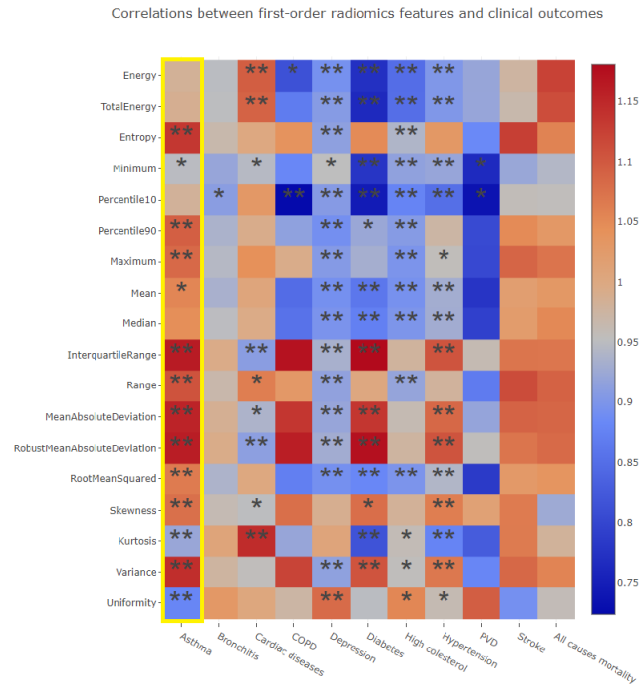


Figure 31: Correlations between first-order radiomics features of the right ventricle in the end of systole phase and clinical outcomes

With this information, I go to the Data showcase section and look for differences in the first-order radiomics features for smokers. With the heatmaps as reference, I can see that the most important features are the entropy, the interquartile range, the mean absolute deviation and the variance. I select these radiomics features for smoking patients, and the results are the following:

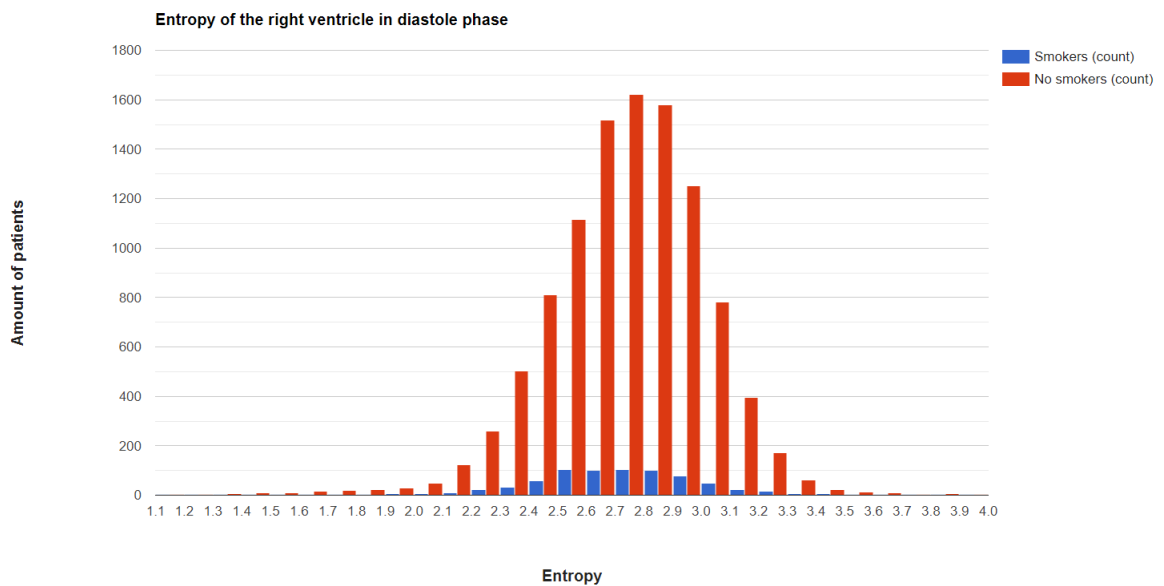


Figure 32: Entropy of the right ventricle in diastole phase for smokers and non smokers

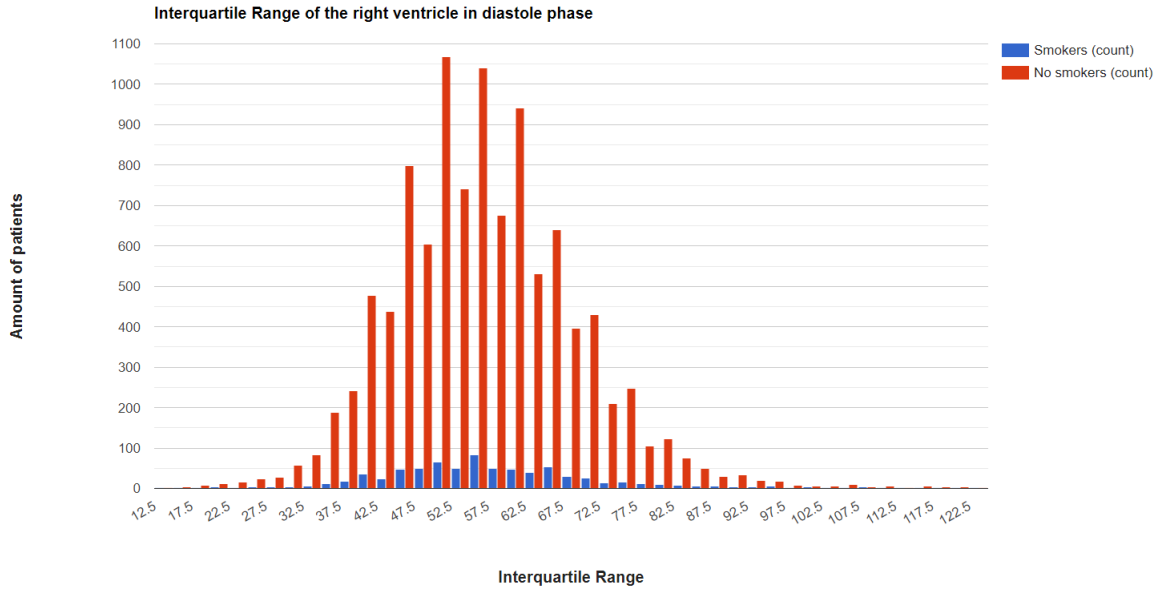


Figure 33: Interquartile range of the right ventricle in diastole phase for smokers and non smokers

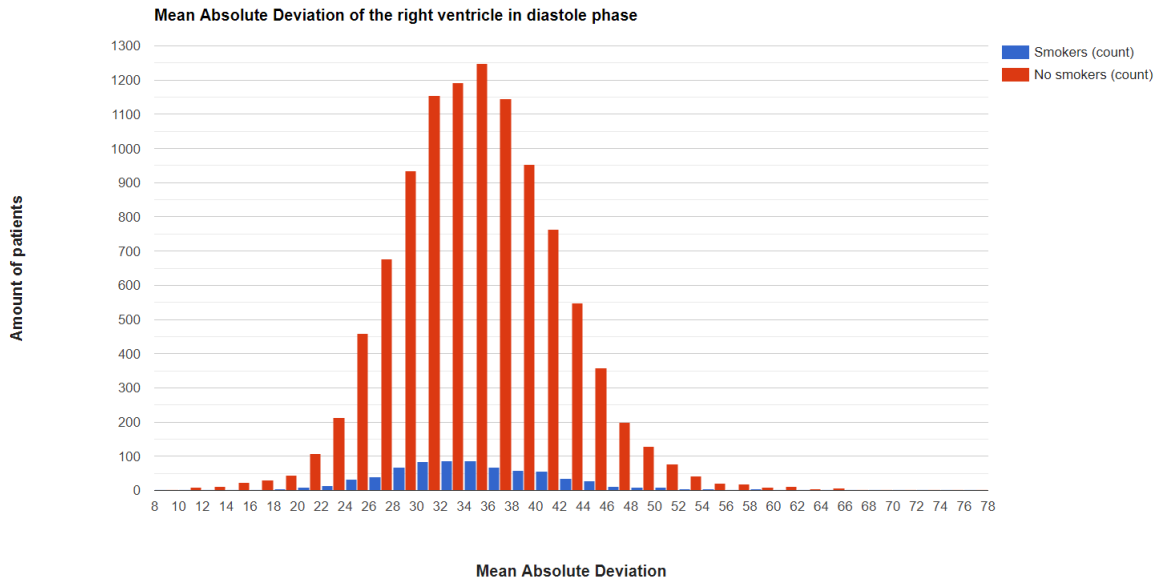


Figure 34: Mean absolute deviation of the right ventricle in diastole phase for smokers and non smokers

The plots correspond to the end of diastole phase, but the differences between these and the plots for the end of systole are negligible. The histograms shows that the mean of the values for these radiomics features for smoking and non smoking patients are very similar, so I can extract the conclusion that smoking does not affect the ratios of getting asthma.

In the study *Factores de desarrollo de asma en la edad adulta*[5] several factors are treated as causes for asthma in adult ages, and smoking is not mentioned.

Specific theories

Now I will use the RWAS web page with more specific purposes. I will formulate hypothesis and try to confirm or refute them using different plots.

Does smoking make the heart smaller?

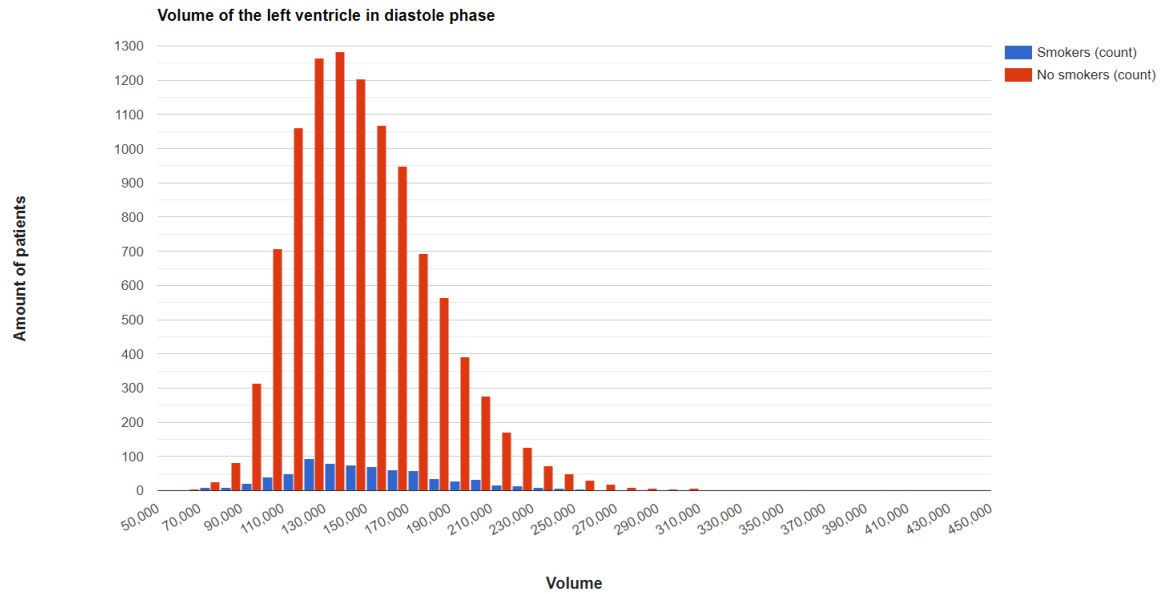


Figure 35: Volume of the left ventricle in diastole phase for smokers and non smokers

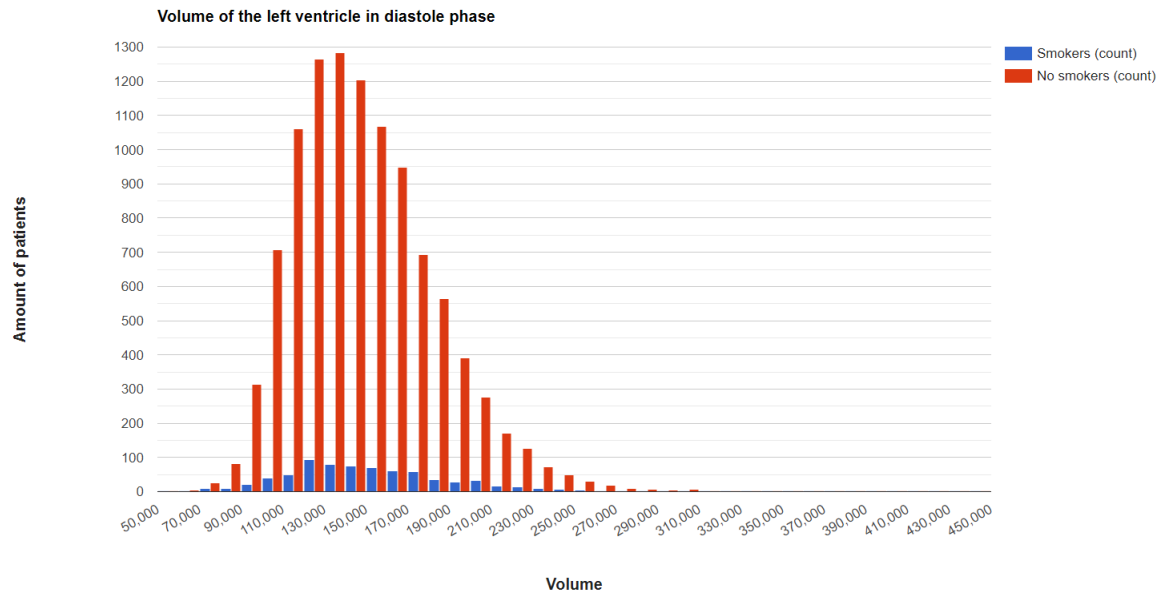


Figure 36: Volume of the right ventricle in diastole phase for smokers and non smokers

The mean of the volume of the heart is almost the same for smokers and non smokers, so smoking doesn't decrease the size of the heart.

Does the size of the heart decrease with the age?

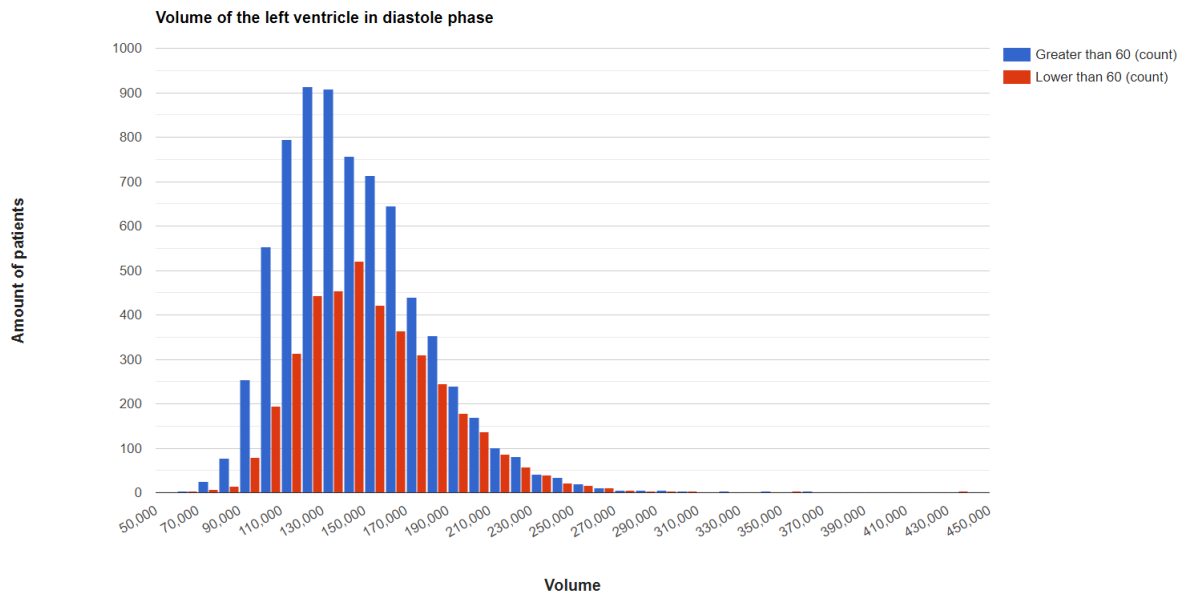


Figure 37: Volume of the left ventricle in diastole phase for patients older and younger than 60 years old

The volume of the heart is smaller for older patients. In conclusion, age is directly correlated with the size of the heart, as it is described in the article *Size*

matters! Impact of age, sex, height, and weight on the normal heart size[6]

Does the heart become less uniform if the patient has high tension?

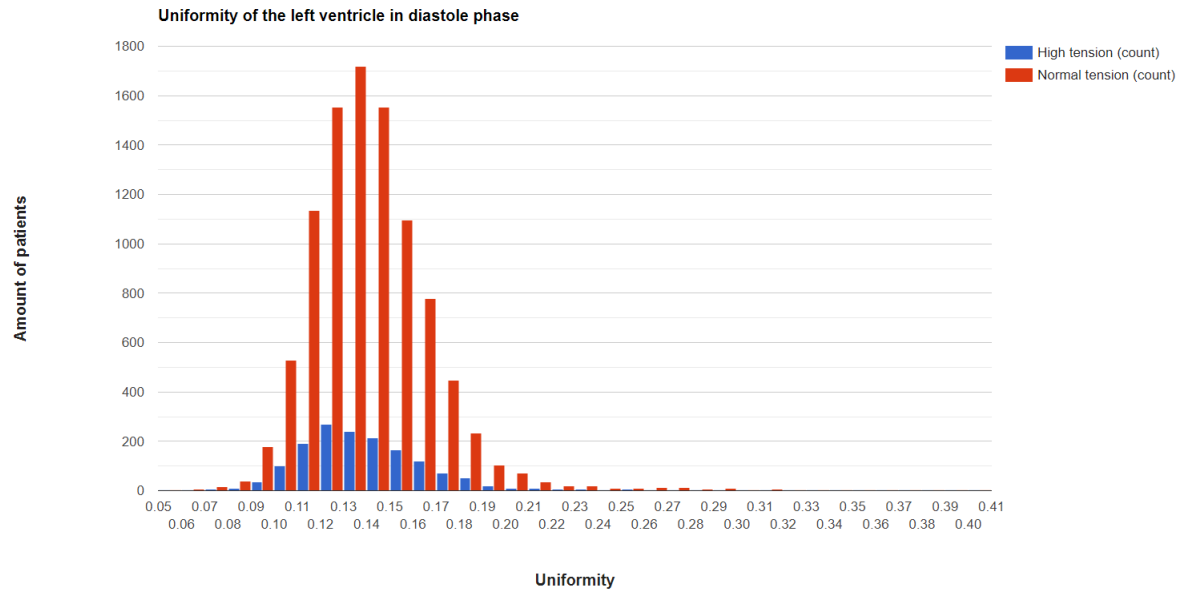


Figure 38: Uniformity of the left ventricle in diastole phase for patients with or without high tension

The left ventricle of patients with high tension is less uniform. This is because high tension damages the vessels.

Is sex a factor on the size of the heart?

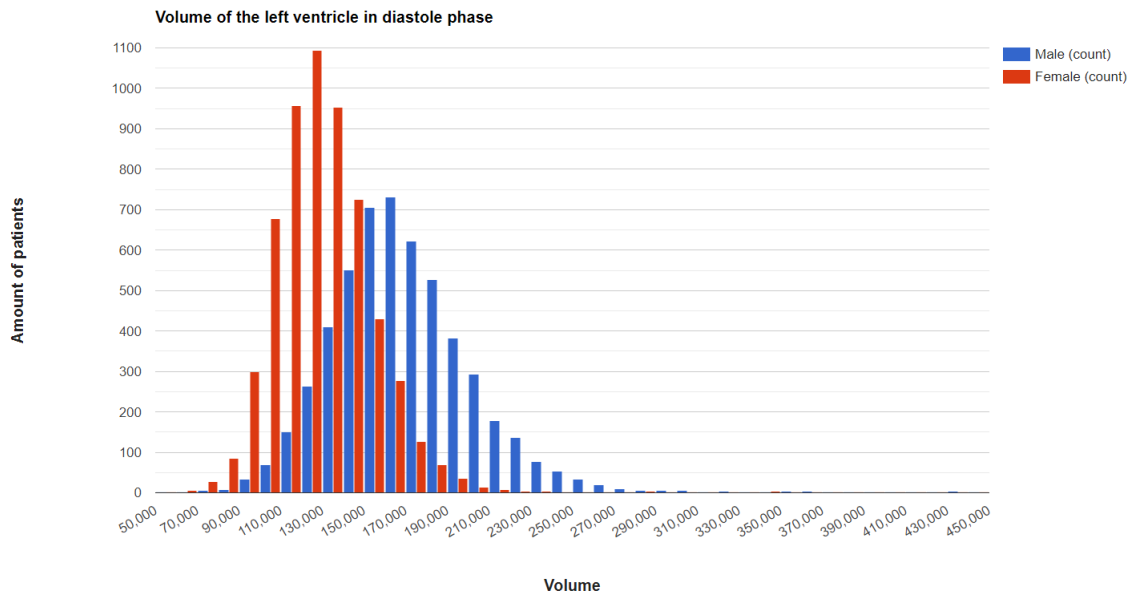


Figure 39: Volume of the left ventricle in diastole phase for males and females

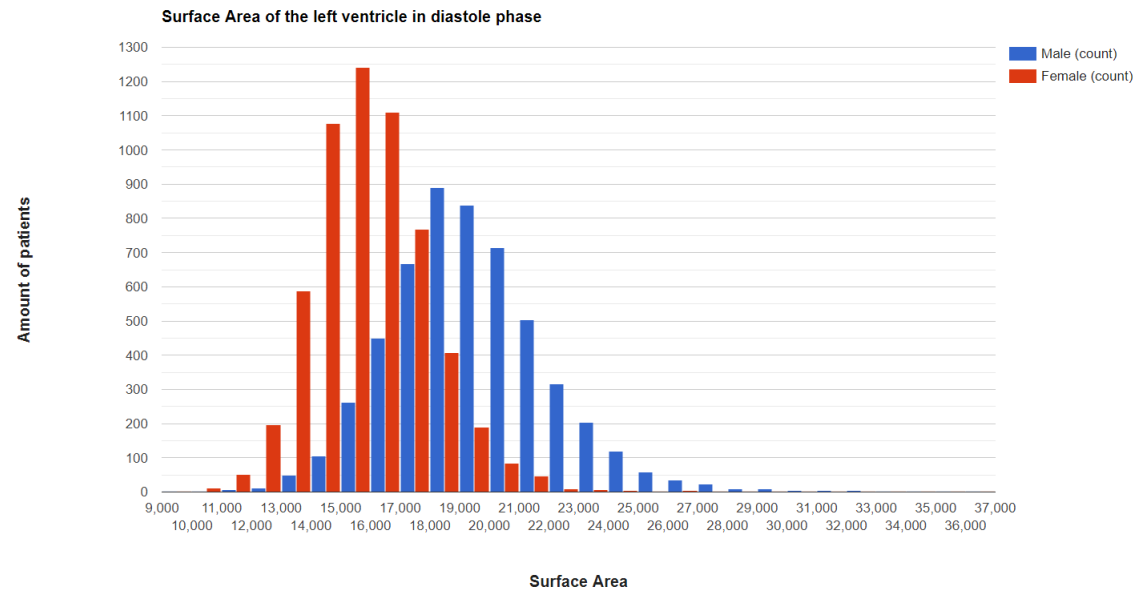


Figure 40: Surface area of the left ventricle in diastole phase for males and females

This demonstrates that males have a bigger heart than females, as it is described in the article *Role of Biological Sex in Normal Cardiac Function and in its Disease Outcome – A Review*[8].

Does asthma alter the shape of the heart?

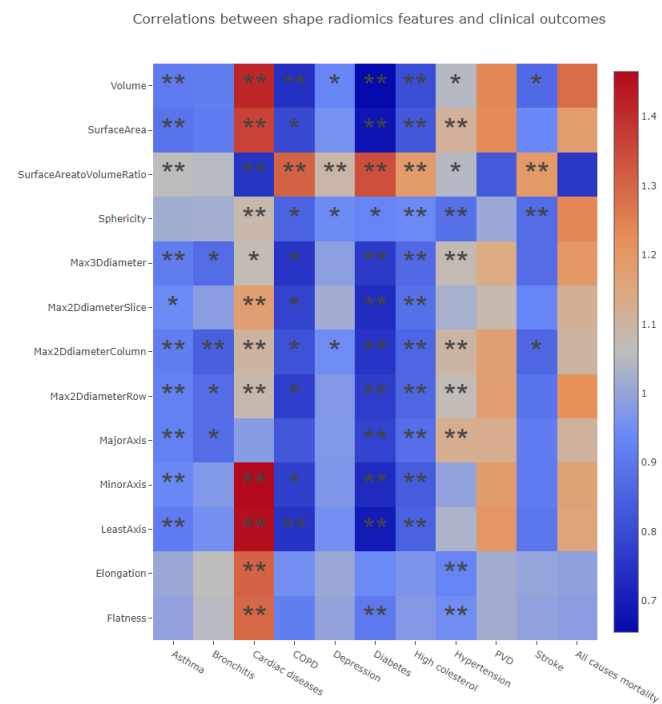


Figure 41: Correlations between shape radiomics features of the left ventricle in diastole phase and clinical outcomes

Asthma is not related to the shape of the heart.

Do respiratory diseases impact the distribution of the heart?

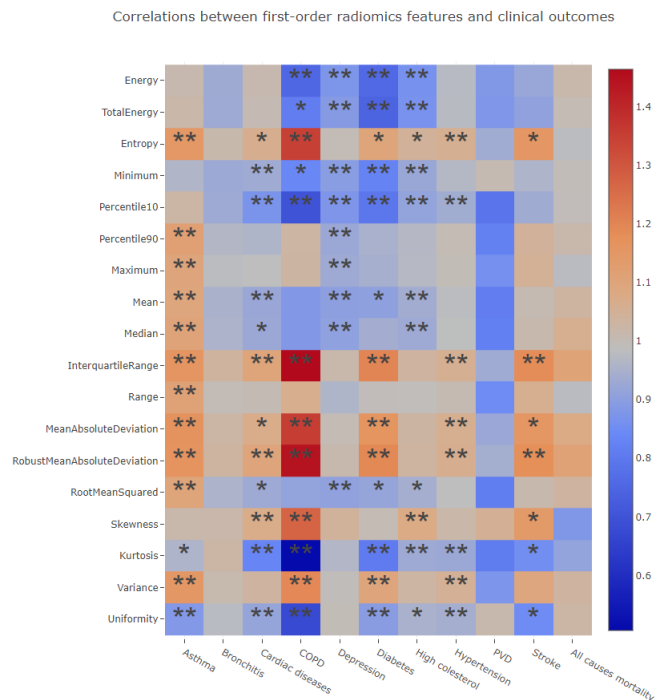


Figure 42: Correlations between first-order radiomics features of the left ventricle in diastole phase and clinical outcomes

I can see that COPD, also known as chronic obstructive pulmonary disease, is correlated with some of the first-order radiomics features, which describe the distribution of the heart. This confirms that respiratory diseases have an impact on the heart.

Is the shape of the heart correlated with mortality?

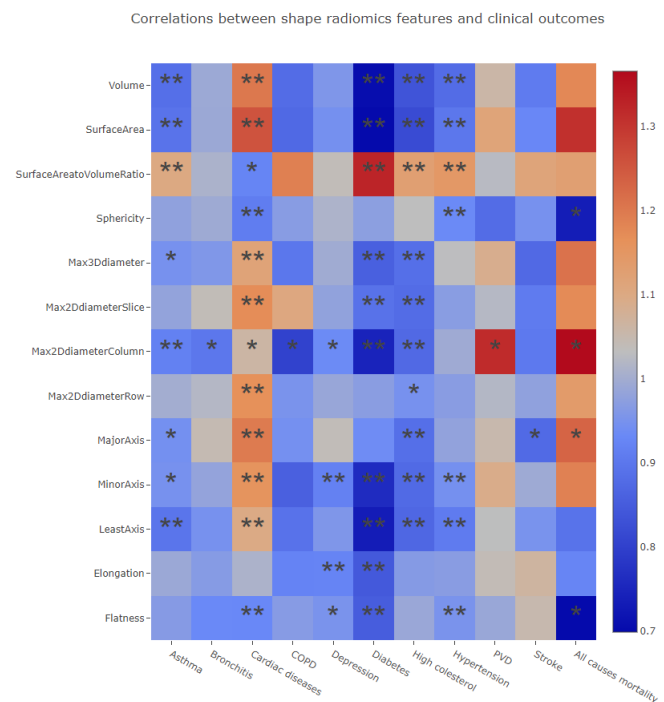


Figure 43: Correlations between first-order radiomics features of the left ventricle in diastole phase and clinical outcomes

The surface area and the maximum diameter of the column of the right ventricle are indicate a risk of mortality.

Does depression impact the heart in any way?

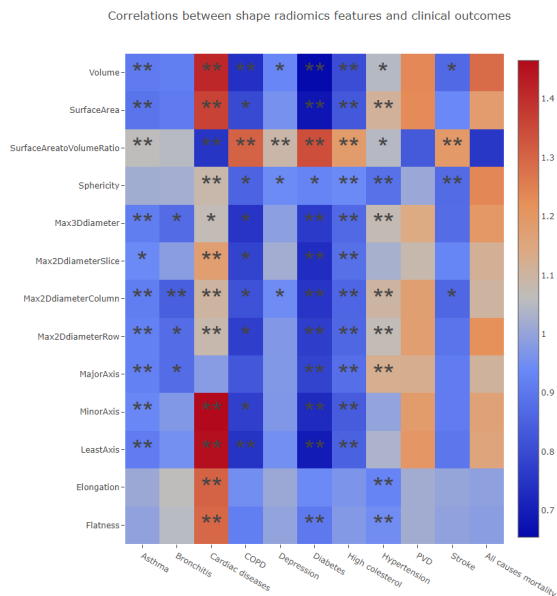


Figure 44: Correlations between shape radiomics features of the left ventricle in diastole phase and clinical outcomes

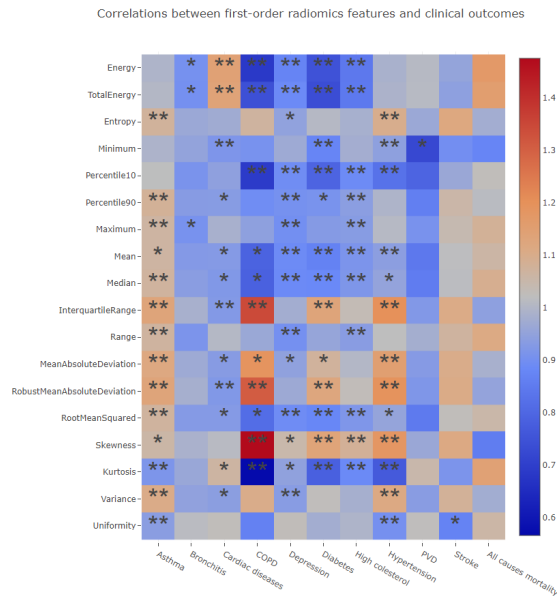


Figure 45: Correlations between first-order radiomics features of the left ventricle in diastole phase and clinical outcomes

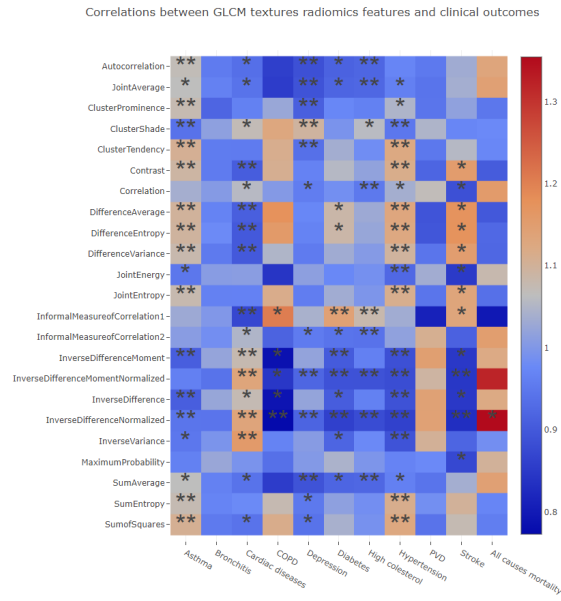


Figure 46: Correlations between GLCM texture radiomics features of the left ventricle in diastole phase and clinical outcomes

These are just some examples of correlations, but the fact is that none of all of them shows a correlation between depression and the heart.

Does a regular alcohol intake affect the heart?

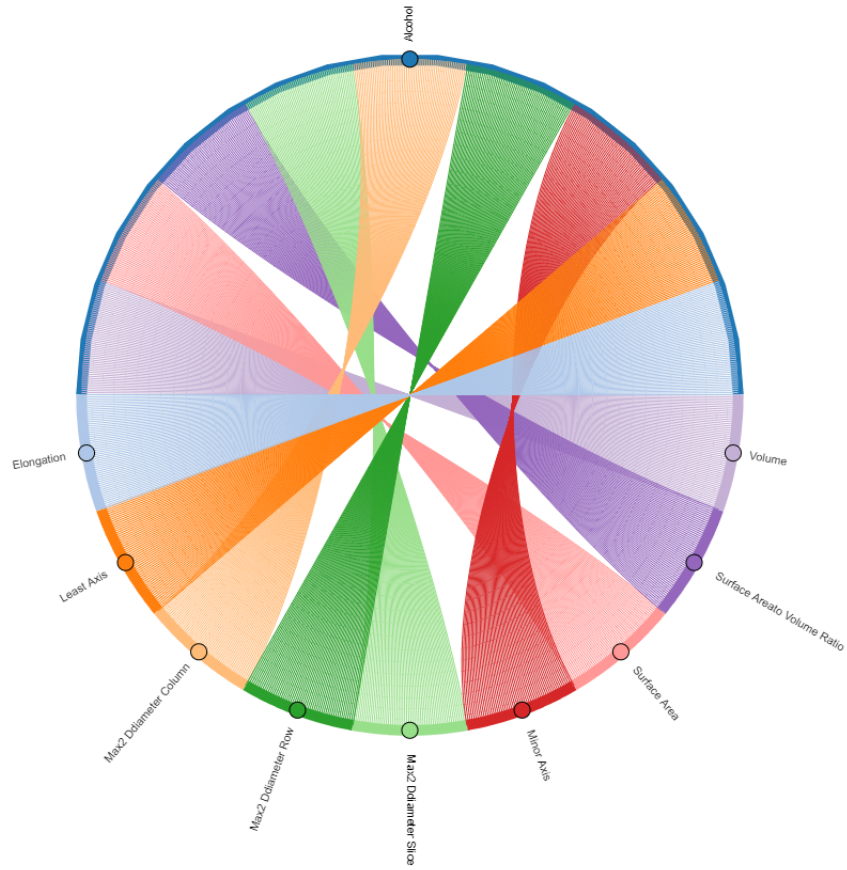


Figure 47: Correlations between shape radiomics features of the left ventricle in diastole phase and cardiovascular risk factors

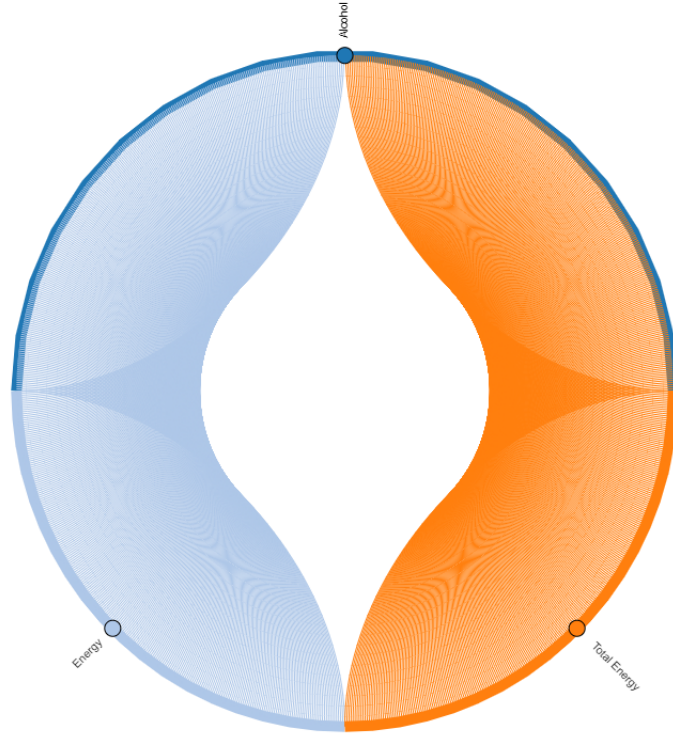


Figure 48: Correlations between first-order radiomics features of the left ventricle in diastole phase and cardiovascular risk factors

Alcohol affects the shape of the heart and it barely affects its distribution. An irregular heart shape can lead to heart diseases, such as an stroke, as described in the article *Alcohol and heart failure*[9].

8 Conclusion

The main objective of developing a useful tool to help investigators formulate hypothesis on the associations between human physical and mental traits and routine behaviours and the odds of getting a disease has been accomplished.

The Radiome wide association study web page satisfies the requirements to be used by anyone with the objective of observing these associations in a very intuitive way. The data has been treated efficiently, making the web site as fast as possible, and the plot types suite every type of data. It also is stylish and user friendly, and still has room for improvement and scaling.

Thinking on the future of the site, three big improvements take the spotlight:

1. Migrating the web page to a private server hosted by the UK-BioBank: as mentioned in a previous section, the RWAS web page is currently being hosted on a free Heroku server. It works fine with the amount of data it uses now, but it would be great to use a more powerful server that allows the web page a larger flow of data and the possibility to scale.
2. Using a proper database: the site gets the data from plain comma separated files and Excel files. Using a secure database as a replacement would increase the speed of the tool and would also make it easier for the amount of data to expand.
3. More data means more plots: finally, the last implementation is simply more plot types. With more plot types, the utility of the RWAS web page would increase and it would allow more variety of uses from investigators.

In conclusion, the Radiome wide association study web page is a good tool for investigators that fulfills the need of visualizing the really the large data the UK-BioBank has on its hands, and it also is a great starting point for a more powerful tool that could improve more significantly the design of health politics.

9 Bibliography

References

- [1] Zahra Raisi-Estabragh, Polyxeni Gkontra, Akshay Jaggi, Jackie Cooper, João Augusto, Anish N Bhuva, Rhodri H Davies, Charlotte H Manisty, James C Moon, Patricia B Munroe, Nicholas C Harvey, Karim Lekadir, Steffen E Petersen *Repeatability of Cardiac Magnetic Resonance Radiomics: A Multi-Centre Multi-Vendor Test-Retest Study*. 2020 Dec 2;7:586236
- [2] Marius E Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, Gary Cook *Introduction to Radiomics*. Apr;61(4):488-495
- [3] Stephen S F Yip, Hugo J W L Aerts *Applications and limitations of radiomics*. Jul 7;61(13):R150-66
- [4] Jaume Roquer, Ana Rodríguez Campello, Meritxell Gomis *Sex differences in first-ever acute stroke*. 2003 Jul;34(7):1581-5
- [5] Urrutia I, Bronte O, Pascual S, Dorado S. *Factores de desarrollo de asma en la edad adulta*. 2018;3(2):46-54
- [6] Stefan Pfaffenberger, Philipp Bartko, Alexandra Graf, Elisabeth Pernicka, Jamil Babayev, Emina Lolic, Diana Bonderman, Helmut Baumgartner, Gerald Maurer, Julia Mascherbauer. *Size matters! Impact of age, sex, height, and weight on the normal heart size*. 2013 Nov;6(6):1073-9
- [7] How high tension impacts the heart. <https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure>
- [8] K. Prabhavathi, K.Tamarai Selvi, K.N. Poornima, A. Sarvanan. *Role of Biological Sex in Normal Cardiac Function and in its Disease Outcome – A Review*. 2014 Aug; 8(8): BE01–BE04.
- [9] Mariann R.PianoRN. *Alcohol and heart failure*. Received 7 November 2001, Revised 13 March 2002, Revised 10 April 2002, Available online 8 February 2003.
- [10] UK-BioBank web site. <https://www.ukbiobank.ac.uk/>
- [11] Flask documentation. <https://flask.palletsprojects.com/en/2.0.x/>
- [12] Google charts documentation. <https://developers.google.com/chart>
- [13] Plotly for javascript documentation. <https://plotly.com/javascript/>
- [14] Holoviews chords plot documentation. <https://holoviews.org/reference/elements/bokeh/Chord.html>