“Exploratory analysis of a biological database (DEXA) and application of Machine Learning models to detect osteoporosis in HIV-positive patients”

Adrià Regué Alsina

Nuria Perez Alvarez
Marc Maceira Duch

Master’s degree in Bioinformatics and Biostatistics UOC-UB
Area 2: Data analysis

January 2021

FITXA DEL TREBALL FINAL

Títol del treball:	*Exploratory analysis of a biological database (DEXA) and application of Machine Learning models to detect osteoporosis in HIV-positive patients*
Nom de l'autor:	*Adrià Regué Alsina*
Nom del consultor/a:	*Nuria Perez Alvarez*
Nom del PRA:	*Marc Maceira Duch*
Data de lliurament:	*01/2021*
Titulació o programa:	*Màster universitari de Bioinformàtica i Bioestadística*
Àrea del Treball Final:	*M0.178 TFM Bioinformàtica i Bioestadística Àrea 2 aula 1*
Idioma del treball:	*Anglès*
Paraules clau:	*HIV, DEXA, Dimensionality reduction, Graphical models, Machine Learning*

Resum del Treball

La incidència d'osteoporosi és major entre la població HIV+, per aquest motiu se'ls realitzen proves DEXA (densitometria òssia) de manera rutinària. Aquest treball s'ha centrat en estudiar una base de dades real, fruit de realitzar aquests anàlisis a pacients amb HIV. Les dades procedeixen de pacients que realitzen el seguiment de la malaltia a la fundació "Lluita contra la SIDA" (Badalona).

És comú que les variables en un estudi mèdic no siguin independents, sinó que estiguin fortament correlacionades. Per això el primer apartat del treball s'ha centrat en purificar la base de dades i descobrir correlacions entre variables mitjançant gràfics de correlacions i mètodes més innovadors com els models gràfics (GGM i MGM). També s'ha aplicat un anàlisi de reducció de la dimensionalitat utilitzant components principals.

En aquest primer punt s'ha corroborat la rellevància que té el gènere en l'estudi. En conseqüència s'ha realitzat tot el treball posterior per cadascun dels gèneres per separat. Els models gràfics apunten a que la importància de

les variables relacionades amb les vèrtebres és trivial a l'hora de calcular el mínim T-score (i per tant, a l'hora de diagnosticar osteoporosi).

La segona part de l'estudi s'ha centrat en generar models predictius capaços de diagnosticar osteoporosis sense utilitzar els marcadors clàssics. S'han aplicat varis algoritmes de Machine Learning (Random Forests, SVM, k-NN) i s'ha generat un model capaç de classificar noves observacions amb una sensibilitat i especificitat del ~80%.

Abstract

Osteoporosis incidence is notoriously larger in the HIV-positive population. For this reason, DEXA analysis (bone densitometry tests) are conducted as a control technique. This work focuses on studying a real DEXA database, retrieved from HIV+ patients doing medical checkups in the *"Lluita contra la SIDA Foundation"*, in Badalona.

Medical databases often suffer from strong correlations between variables. For this reason, the first chapter of the study has been destinated to purify the database and discover said relationships, via correlation plots and more innovative techniques such as graphical models (GGMs and MGMs). Also, a dimensionality reduction analysis has been executed using principal components.

This first part of the study corroborated the relevance of the gender variable. All the subsequent analysis has been conducted separately for men and women. Also, graphical models suggested that vertebral variables have a rather weak importance in determining the minimum T-score (and thus, predicting osteoporosis).

The second part of the study has focused on generating a predictive model with the ability to diagnose osteoporosis without using its classical indicator variables. After modelling with various Machine Learning algorithms (Random Forests, SVMs, k-NNs), a classificatory model has been generated, reporting a sensitivity and specificity of ~80%.

Index

List of figures and tables

1. Introduction

1.1 HIV and common comorbidities

HIV is a worldwide-spread viral disease that affects the functionality of the immune system by destructing leucocytes. Therefore, infected patients are unable to activate an immune response and are more prone to suffer other viral/bacterial infections along with some types of cancers. The most advanced stage of the disease is known as AIDS, although not all HIV-infected patients will develop it.

Nowadays there is no healing treatment for HIV, and the medical approaches focus on decreasing the viral load and the reproduction rates. The chemicals used to achieve those lines of therapy are known as antiretrovirals (ARVs) and are classified in 7 groups depending on their molecular targets.

In 2019, 1,7 million people got infected and 700.000 died of HIV-related diseases, according to UNAIDS (UNAIDS, 2020). The latest estimations estate that there are 38 million people living with the infection, of which only 25,4 (about a 66%) are following an ARV treatment.

The knowledge around the disease and its possible treatments has seen an exponential increase in the last decades. Early detection and a medical follow-up, along with the discovery of new treatments, has supposed a big improvement in both the quality of life and longevity of the patients. In 2019, the deaths by HIV-related diseases were roughly a 60% of the number of deaths in 2010 (UNAIDS, 2020).

This improvement in life longevity lead doctors and scientists to discover a number of comorbidity disorders associated with the ageing of HIV-infected patients. Recent studies (Finnerty, Walker-Bone, & Tariq, 2017; Negredo et al., 2018) infer a direct relation with the presence of the disease and various lean, fat and bone mass anomalies, likely caused by the constant inflammation of the tissues and as a side effect of the ARV drugs (Compston, 2016).

Patients with low bone density have a higher risk of suffering from vertebral, hip bone and limb fractures (Premaor & Compston, 2020). Women aged 45 to 56 following some ARV treatment seem to be a high-risk population, because of the decrease in the estrogen levels due to menopause.

Also, there seems to be a positive relationship between low bone density (either osteopenia or osteoporosis) and the use of antiretroviral drugs: patients lose between a 2 and a 6% of bone mineral density (BMD) in the first years of treatment (Compston, 2016).

Appearance of muscular dystrophy is 1,1 to 33,5 times higher in patients with HIV (Oliveira, Borsari, Webel, Erlandson, & Deminice, 2020), but HIV is still not considered a risk factor.

ARV treatment has also been linked to lipodystrophy. Patients usually show lipidic accumulation in abdomen, chest, and neck (lipohypertrophy), while showing lipidic loss in face, limbs, and waist (lipoatrophy). It is also common the appearance of lipomas (Guzman & Vijayan, 2020).

1.2 State of the art and study justification

Health related databases usually contain mixed data (categoric, numeric…) of uncommon origins (routine controls, medical prescriptions…) and great complexity. Also, number of variables tend to be large respect the number of observations (i.e., a blood test from a single patient contains over 80 variables analyzed). Therefore, preparation, analysis and global interpretation of the data is a complex duty and requires modern algorithms (usually based in machine learning) to be achieved.

When it comes to interpreting results of DEXA analysis of HIV-positive patients, as is the case in this study, data is usually looked over and only the stablished marker variables are considered to diagnose a series of comorbidities in the patients (i.e., minTscore for osteoporosis, appendicular lean mass for LMM, etc).

This study will be conducted under the hypothesis that a deeper exploratory analysis can be useful to further understand the relationships between variables, along with detecting possible dynamics that are overlooked in the classical studies.

Prior studies tried to deepen in the database complexity via dimensionality reduction (PCA) and regression/classification modeling with Random Forest approaches.

The scientific community has been strongly focused in discovering new algorithms of increased complexity, that could study conditional correlations between variables of different natures (as is the case in this study). Recently, a method has been reported (Altenbuchinger, Weihs, Quackenbush, Grabe, & Zacharias, 2020; Sedgewick et al., 2019) claiming to be able to interpret multimodal data and thus opening some boundaries in the medical-statistical field. In this study, the data will be analyzed under some ML methods, including the vanguardist Gaussian Graphical Models and Mixed Graphical Models, aiming to discover some relationships hitherto unknown.

Results are expected to help understand the synergies between variables and the comorbidity markers, opening a door to more accurate disease diagnose and detection.

1.3 Main objectives

Three main objectives are expected to be covered in the extension of this study. They are:

1. **Study the complexity of the database, the correlations between variables using non-supervised models and perform a dimensionality reduction approach.**
    a. Determine how missing data, typos, repeated values and outliers have to be treated.
    b. Univariate analysis of the dataset: descriptive statistics.
    c. Establish the underlying relationships and correlations between the variables, via correlation plots and graphical models.
    d. Determine the importance of the gender variable.
    e. Dimensionality reduction via Principal Component Analysis.
2. **Design of a predictive model for osteoporosis using the variables of the database.**
    a. Study of the modeling options available for our type of variables and application of the best model/s.
    b. Application of regression and classification models over the osteoporosis variables (ampliation work: model LMM and lipodystrophy).
    c. Selection of the model with best performance.
3. **Generate a dynamic report in Rmarkdown format, that allows not only to replicate the analysis but to perform it over a new dataset.**

1.4 Work line

A deep browse in various search engines (Pubmed, Scopus, Web of Science…) has been conducted to determine the State of the Art about statistical methods. Most authors address their studies conducting an exploratory analysis, that leads to a directed analysis to obtain results to their hypothesis. This is the structure that has been followed in this study.

The first step in this study has been to load the data into the statistical software and transform it for the exploratory analysis. This initial step focused on addressing missing values, outliers, possible typos, merge databases, calculate missing variables, etc. The objective has been to have a consistent database to perform the exploratory analysis, which is the second step. Also, one of the major

concerns has been solved in this first step, which was whether genders had to be treated as a single group or separately.

In the exploratory analysis a descriptive study of our variables has been performed, studying their normality, distribution, linear correlations, etc. A dimensionality reduction approach has been conducted, using Principal Components Analysis. This method has been selected for being the most used in literature.

Relationship between variables has been furtherly studied using both directed and undirected graphical models.

Finally, many machine learning algorithms have been selected to infer sample classification and regression based on subsets of the database variables. Random Forests, Support Vector Machines and k-Nearest Neighbors have been performed. Classical models such as linear and logistic regressions have been omitted in favor of the more powerful methods.

1.5 Workplan

The first step of this project is reading the state of the art, followed by data incorporation and processing. Once the general context has been established, future work has been divided in three main steps: 1) Non-supervised analysis; 2) Statistical models for osteoporosis and 3) formatting the Rmakrdown document.

The first and second objectives have been designed as follows:

1. **Non-supervised analysis** **14 days**
    a. Bibliographic research about the state of the art 7 days
    b. Data filtering and processing 2 days
    c. Descriptive statistics 2 days
    d. Correlation plots 2 days
    e. Bibliographic research about unsupervised methods 2 days
    f. Unsupervised analysis 7 days
    g. Gender effect study 7 days
    h. Methods and results annotation 9 days

2. **Statistical models for osteoporosis** **42 days**
    a. Bibliographic research about the best models 7 days
    b. Database division in train and test 1 day
    c. Model design and application over osteoporosis 20 days
    d. Methods and results annotation 15 days
    e. Models over LMM and lipodystrophy (ampliation) 20 days
    f. Conclusions 20 days

The third objective (**3. Dynamic report**) has been designed to be performed side by side with the 2.f point, with a length of 20 days.

When the statistical part of the project ended a last milestone started: the redaction of the paper.

4. Paper redaction	**41 days**
a. Choosing a copyright license	1 day
b. State of the art	3 days
c. Introduction and background	8 days
d. Materials and methods	8 days
e. Results	8 days
f. Conclusions	16 days
g. Abstract	16 days
h. Bibliography management	8 days
i. Annexes	8 days
j. Glossary	8 days
k. Figure and images translation	8 days
l. Last reading	13 days
m. Visual presentation	17 days

A Gantt graphic of the detailed milestones and the corresponding partial evaluations can be seen in the Figure 1. In addition to the upper information, the following image contains the monitoring reports and the PACs (Continuous Evaluation Tests).

Weak colors inside a group indicate that said milestones are designed as ampliation work. Completion of additional milestones is strongly dependent to the availability of time.

Modeling is the most delicate part of the study, and the one that is more likely to suffer from time deviances. For this reason, studying the other two comorbidities has been designed as ampliation work.

There are two determining moments in the calendar: the 25th of October (ending of the non-supervised analysis and starting of the data modeling) and the 1st of December (paper redaction, which requires the modeling to be completed).

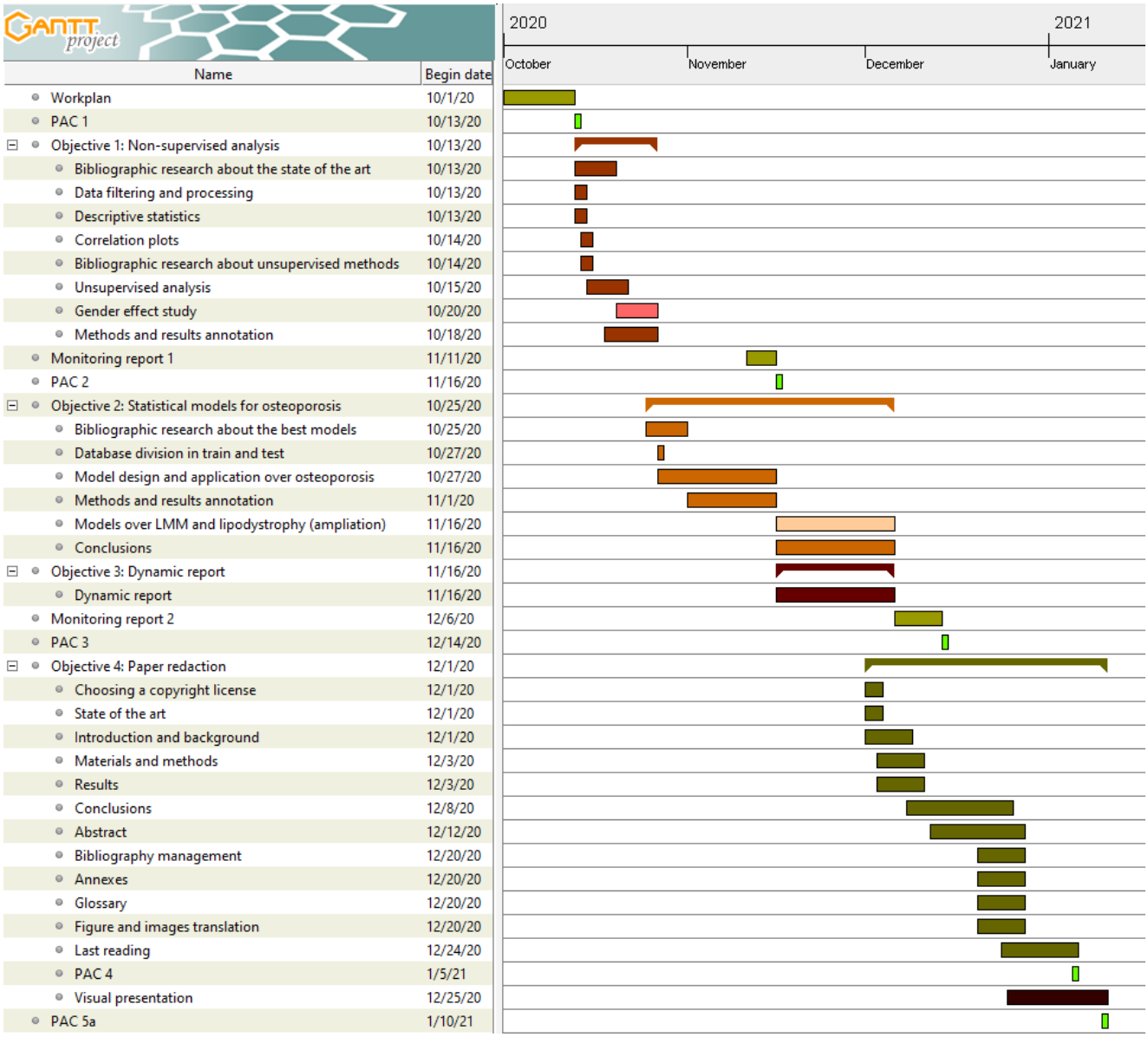


Figure 1: Gantt graphic. Diagram detailing the objectives and milestones followed in the project, generated using the freeware “GanttProject”.

1.6 Final products

This final project retrieves a ML based model to predict osteopenia/osteoporosis presence in HIV patients based on their muscular and fat values. It also increases the knowledge about the relationships that exist between muscular and bone-related variables, along with a practical application of vanguard methods such as the Gaussian and Mixed graphical models.

As a side product, a dynamic report is generated (Rmarkdown format), to replicate the statistical study over the same data or over new observations.

1.7 Chapter description

Chapter 2: The "DEXA" database gives some background about the database and its variables. This chapter is designed to introduce some biological concepts to the reader and provide him with all the technical information to follow the work done in the study. Along with the theorical concepts, the origin of the database and its nature are explained, focusing on the different comorbidities and the way they are calculated with the variables of the database.

Chapter 3: Statistical background will introduce the reader to the concepts of graphical models, machine learning, random forests, support vector machines and k-Nearest neighbor algorithms. This chapter gives a theorical background about the models used in the study.

The following two chapters, **4: Materials and methods** and **5: Results** are designed to follow the structure of any scientific paper. The first chapter contains the information of the steps and analysis taken in the study, while the later returns the results of said operations. Both chapters follow a side-by-side structure, starting by the univariate study and ending with the directed analysis.

Chapter 6: Conclusions closes the technical part of the study, summing up the most relevant information that can be obtained of the previous chapters. All the hypothesis and products obtained are detailed in this chapter. Not only the study-related conclusions are included in this section, but also a generic view about the work done, timelines and personal growth is provided.

Chapter 7: Glossary gives brief descriptions of the most relevant terms of the study. I focused on including the statistical tests (such as the T-test, or Shapiro-Wilk), statistical models (PCA, random forest, SVM…) and performance-related terms (sensitivity, specificity, kappa…).

Chapter 8: Bibliography contains the relation of all the books and papers used to retrieve information. They have been indexed following the American Psychological Association (6th edition) citation style.

Chapter 9: Annexes is a section dedicated to all those materials that are not required to follow the study but provide extra information and/or transparency to the obtained results. Summary tables and graphics excluded from the main text can be found in the annexes (also, graphics excluded from the annexes can be retrieved by contacting the author).

2. The “DEXA” database

2.1 Origin of the data and variables

Osteoporosis is one of the most well-established comorbidities associated with HIV and its detection and monitoring is highly protocolized. For this reason, patients with HIV usually are also under control for bone-related diseases and are periodically exposed to a Dual-energy X-ray absorptiometry (DEXA).

DEXA is an imaging technique that uses radiation to measure the bone mineral density in various body zones. Several bones are measured in every screening, usually vertebrae (T1 to T4), hip bone and femur (trochanter, neck and Walds). Muscular and fat values are also retrieved in the same scan.

The data used in this study has been collected in the *“Fundació Lluita contra la SIDA”* in the Germans Trias i Pujol University Hospital, Badalona. It reflects the DEXA analysis conducted on HIV-infected patients in the last 20 years. While some of these screenings are conducted under medical prescription, others reflect routine controls. This represents a challenge when it comes to analyzing the data, as both groups may differ in behavior and can bias our study.

The DEXA analysis summarizes a series of body measures, that can be classified as bone-related, lean mass-related and fat-related. In the next table (Table 1) a comprehensive list of all variables in the database is provided, colored according to said classification (variables in orange refer to the lean and fat mass values, while variables in green are bone related). If any transformation improves their normality it has been listed in the description.

Variable name	Description	NAs	Origin
ID	Patient identification number. Factor.		DB 1*
gender	Gender of the patient. Factor: 2 levels (“F”, “M”).		DB 1
gender_num	Gender coded as binary. Factor: 2 levels (0 = F, 1 = M).		DB 1
Age	Age, in years, at the time of the DEXA. Numeric.		DB 1
Age_cat	Patient classification in under/over 50 years old. Factor: 2 levels (<=50, >50).		DB 1
Height	Full body height, in meters. Numeric.	2	DB 1
Weight	Full body weight, in Kg. Numeric.	1	DB 1
RAFp	Right arm fat, percentage. Numeric.		DB 1
RAFg	Right arm fat, grams. Numeric. Log transform.		DB 1
RALg	Right arm lean, grams. Numeric. Log transform.		DB 1
LAFp	Left arm fat, percentage. Numeric.		DB 1
LAFg	Left arm fat, grams. Numeric. Log transform.		DB 1
LALg	Left arm lean, grams. Numeric. Log transform.		DB 1
BothAFp	Both arms fat, percentage. Numeric.		DB 1
BothAFg	Both arms fat, grams. Numeric. Log transform.		DB 1
BothALg	Both arms lean, grams. Numeric. Log transform.		DB 1

RLFp	Right leg fat, percentage. Numeric.		DB 1
RLFg	Right leg fat, grams. Numeric. Log transform.		DB 1
RLLg	Right leg lean, grams. Numeric. Log transform.		DB 1
LLFp	Left leg fat, percentage. Numeric.		DB 1
LLFg	Left leg fat, grams. Numeric. Log transform.		DB 1
LLLg	Left leg lean, grams. Numeric. Log transform.		DB 1
BothLFp	Both legs fat, percentage. Numeric.		DB 1
BothLFg	Both legs fat, grams. Numeric. Log transform.		DB 1
BothLLg	Both legs lean, grams. Numeric. Log transform.		DB 1
TFp	Trunk fat, percentage. Numeric.		DB 1
TFg	Trunk fat, grams. Numeric. Square root transform.		DB 1
TLg	Trunk lean, grams. Numeric. Square root transform.		DB 1
TotalFp	Total body (no head) fat, percentage. Numeric.		DB 1
TotalFg	Total body (no head) fat, grams. Numeric. Square root transform.		DB 1
TotalLg	Total body (no head) lean, grams. Numeric. Log transform.		DB 1
L1BMD	Vertebra 1, BMD value. Numeric.	1	DB 1
L1T	Vertebra 1, T-score value. Numeric.	1	DB 1
L1Z	Vertebra 1, Z-score value. Numeric.	1	DB 1
L2BMD	Vertebra 2, BMD value. Numeric.	1	DB 1
L2T	Vertebra 2, T-score value. Numeric.	2	DB 1
L2Z	Vertebra 2, Z-score value. Numeric.	1	DB 1
L3BMD	Vertebra 3, BMD value. Numeric.	1	DB 1
L3T	Vertebra 3, T-score value. Numeric.	2	DB 1
L3Z	Vertebra 3, Z-score value. Numeric.	2	DB 1
L4BMD	Vertebra 4, BMD value. Numeric.	2	DB 1
L4T	Vertebra 4, T-score value. Numeric.	2	DB 1
L4Z	Vertebra 4, Z-score value. Numeric.	3	DB 1
L1L4BMD	Vertebrae 1-4, BMD value. Numeric.	1	DB 1
L1L4T	Vertebrae 1-4, T-score value. Numeric.	1	DB 1
L1L4Z	Vertebrae 1-4, Z-score value. Numeric.	1	DB 1
L2L4BMD	Vertebrae 2-4, BMD value. Numeric.	2	DB 1
L2L4T	Vertebrae 2-4, T-score value. Numeric.	2	DB 1
L2L4Z	Vertebrae 2-4, Z-score value. Numeric.	2	DB 1
NeckFBMD	Femoral neck, BMD value. Numeric.		DB 1
NeckFT	Femoral neck, T-score value. Numeric.		DB 1
NeckFZ	Femoral neck, Z-score value. Numeric.		DB 1
WardsBMD	Wards region, BMD value. Numeric.	1	DB 1
WardsT	Wards region, T-score value. Numeric.	1	DB 1
WardsZ	Wards region, Z-score value. Numeric.	1	DB 1
TrochBMD	Greater trochanter, BMD value. Numeric.	1	DB 1
TrochT	Greater trochanter, T-score value. Numeric.	1	DB 1
TrochZ	Greater trochanter, Z-score value. Numeric.	2	DB 1
TotalFBMD	Total femoral values, BMD value. Numeric.		DB 1
TotalFT	Total femoral values, T-score value. Numeric.		DB 1

TotalFZ	Total femoral values, Z-score value. Numeric.		DB 1
BMI	Body mass index, total Kg/m^2. Numeric. Log transform.	3	DB 1
BMI_cat	BMI categorical. Factor: 4 levels (“Underweight”, “Normal”, “Overweight”, “Obesity”)		DB 1
FMI	Fat mass index, Fat Kg/m^2. Numeric. Log transform.	3	DB 1
FFMI	Fat free mass index, lean Kg/m^2. Numeric. 1/x transform.	3	DB 1
Apendicularleanmas	Appendicular lean mass index, Kg/m^2. Numeric.	2	Calculated
FMR	Fat mass ratio: trunk fat % / legs fat %. Numeric. Log transform.		DB 1
FTrunkgFLegsg	Trunk fat mass / legs fat mass. Numeric. Log transform.		Calculated
Indexdistributionfat	Trunk fat mass / all limbs fat mass. Numeric. Log transform.		DB 1
FtrunkpFlimbsp	Trunk to limbs ratio: trunk fat % / limbs fat %. Numeric. 1/x transform.		DB 1
FtrunkgFtotalg	Trunk fat mass / total fat mass. Numeric.		DB 1
FLegsgFtotalg	Legs fat mass / total fat mass.		DB 1
FlimbsgFtotalg	Limbs fat mass / total fat mass. Numeric.		DB 1
LLegFgBMI	Left leg fat mass / BMI. Numeric.	2	DB 1
LLegFpBMI	Left leg fat percentage / BMI. Numeric.	2	DB 1
Lipodystrophy	Presence of lipodystrophy. Factor: 2 levels (0, 1=presence).		
Sarcopenia	Presence of sarcopenia. Factor: 2 levels (0, 1=presence).	2	
LipoSarcop	Presence of lipodystrophy and/or sarcopenia. Factor: 2 levels (0, 1=presence).	2	
phenotype	Combination of outcomes between BMI_cat and lipodystrophy/sarcopenia. Factor: 4x2x2 = 16 levels.	2	
minTscore	Lowest T-score value observed. Numeric.		Calculated
Tscore_3cat	WHO classification of bone density loss. Factor: 3 levels (“Healthy”, “Osteopenia”, “Osteoporosis”).		Calculated
TotalBMD	Total body BMD. Numeric.	12	DB 1
HIV_date	HIV diagnose date. POSIX time value.	28	DB 2**
dexa_date	Date of the DEXA screening. POSIX time value.	18	DB 2
Disease_age	Time passed between HIV diagnosis and DEXA, in years. Numeric.	28	Calculated

Table 1: Variables in the databases. Columns represent the name of the variable in the database, a brief description, the number of missing values and the origin (Database 1, 2 or manually calculated).
***** *DB 1: “LastDEXA_10-05-18English_selection2 (1) (1).sav”, renamed as “data DEXA.sav”.*
****** *DB 2: “DEXA amb demogràfics_Vitruvi_13-01-20.xlsx”, renamed as “DEXA_nova.xlsx”.*

2.2 The “minTscore” and “Tscore_3cat” variables

Bone mineral density is a measure of the healthiness of the bone and an indicator for various bone-related diseases (i.e., osteoporosis and osteopenia). BMD is measured as grams of calcium hydroxyapatite per cm^3, but it is almost never used in its form. Instead, two statistical parameters are calculated: T-scores and Z-scores.

Both the T and Z-scores represent how many standard deviations a given BMD is away from the population value. The difference between the two statistics

resides in the population which the value is compared against: while the T-score is compared against a healthy male adult of 30 years old, the Z-score is computed against a healthy population of the same age and gender as the patient.

In the practice T-score is used in most situations, Z-score having a role only with children and pre-menopausal women.

The T-score values determine the level of demineralization and are a common cutting value to diagnose osteoporosis and osteopenia (Liu et al., 2011; Oursler et al., 2020). The World Health Organization establishes the following criteria:

T-score above -1:	Healthiness
T-score between -2.5 and -1:	Osteopenia
T-score below -2.5:	Osteoporosis

Assuming that BMD follows a Gaussian distribution, these values would roughly represent the 84% (-1) and 99.4% (-2.5) percentiles.

The bone mineral density is measured at a few locations. L1 to L4 represent individual vertebral values; L1L4 and L2L4 are mean values for the groups of vertebrae. The femoral measurements are taken at different parts: greater trochanter, Wards region, femoral neck, and total femoral bone.

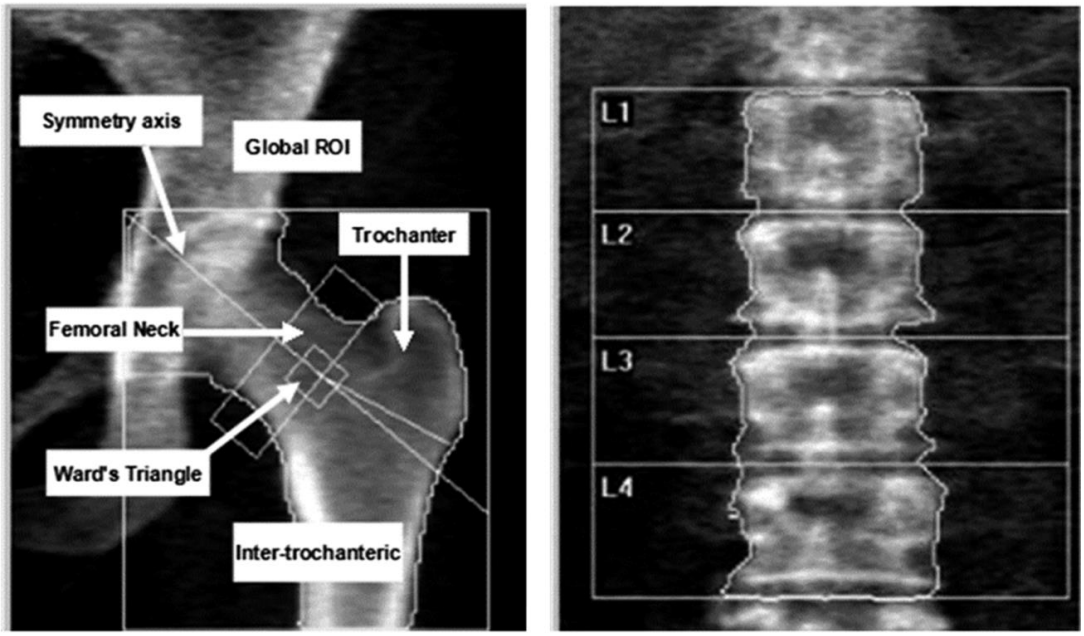


Figure 2: Delimitation of the bones in a DEXA analysis. Left: thigh bone zones: trochanter, femoral neck and Ward's triangle. Right: vertebrae classification. *(Font: Doroudinia & Colletti, 2015)*.

The "minTscore" variable represents the minimum of the T-scores of every patient. This value is used to calculate the "Tscore_3cat", which classifies every patient in one of the three categories (healthy, osteopenia, osteoporosis) based on the criteria mentioned above.

Both BMD and T/Z-scores are strongly dependent of the placement of the patient inside the DEXA scanner and thus we rely on the ability of the screening technician. Spinal deformities are also a common handicap for vertebrae BMD lectures.

2.3 The “lipodystrophy” variable

Lipodystrophy is a pathology characterized by an abnormal redistribution of fat tissue along the body, that can be expressed by loss and/or gain of lipidic mass. It is common that patients show a decrease in the facial, appendicular and backside fat, while at the same time accumulate higher levels of lipidic tissue it in the abdominal/pectoral region and the neck. The symptoms usually appear as a consequence of the antiretroviral treatment and are associated with decreased self-esteem and depression, that can eventually lead the patients to leave the drug treatment (Guzman & Vijayan, 2020).

Beyond the psychological problems associated with the pathology some physical complications are observed, such as insulin resistance, cardiovascular diseases and lipomas (Guzman & Vijayan, 2020).

2.4 The “sarcopenia” variable

Sarcopenia is a disorder characterized by a loss of both muscular tissue and muscle function. It affects all the muscular tissue of the body and degenerates over time. Patients with this disorder have an increased likelihood of falling and suffering fractures.

There are several known causes for sarcopenia, including ageing, lack of physical activity, changes in hormones and malnutrition. However, it has been seen that patients living with HIV have an increased risk of suffering from muscular weakness.

Although we use the name “sarcopenia” in the dataset, it would be more accurate if we talked about low muscle mass, since we will be making our decisions based on the appendicular lean mass values retrieved by the DEXA (muscular mass in relation to height). A diagnosis of sarcopenia needs to be associated with muscular weakness and requires a strength test (Studenski et al., 2014).

3. Statistical background

This chapter is designed to provide some theorical background about the selected graphical and machine learning models. We will talk about their method of action, along with their strengths, weaknesses and limitations.

Graphical Models

It is common in biological datasets to find out that data is very intercorrelated. This problem has been studied via classical approaches (i.e., pairwise correlation matrices and plots), but said methods are very weak in the sense that they do not distinguish direct relationships from dependencies mediated by a third variable. A more recent approach to face the correlation problem is using full order partial correlations, i.e., correlation between two variables accounting for all the other variables in the dataset.

We say that two variables are independent if the values that they take are not influenced one-another. If we have more than two variables, we must consider the possibility that the (in)dependence between two of them is mediated by a third one. As an extension, two variables are conditionally independent of a third one if the values they take are not influenced by the values of the third variable. Expressed mathematically: $P(A=a \cap B=b \mid C=c) = P(A=a|C=c)P(B=b|C=c)$.

Probabilistic graphical models (PGMs) are algorithms that visually represent these full order partial correlations, using nodes to represent the variables and edges to symbolize the dependencies. The models that only represent numerical, normally distributed data are known as Gaussian Graphical Models (GGMs).

While GGMs are the simplest of the models and are easy to interpret, they tend to overfit relationships. An approach to reduce the number of edges is using the LASSO (Least Absolute Shrinkage and Selection Operator) regularization method, which assumes that most of the possible edges in the graphic are equal to zero. LASSO applies a penalty parameter, generating simpler graphics that are easier to interpret and are more prone to represent the true relationships between variables.

The value of this shrinkage parameter is unknown and needs to be determined statistically. Usually, cross-validation or EBIC strategies are used.

If the assumption of normally distributed data is violated, or if we want to incorporate factorial variables in the study, then a different method must be used. Mixed Graphical Models (MGMs) are a novel method that allows us to incorporate such variables.

Machine Learning

Professor Tom M. Mitchell defined ML as *"[…] the study of computer algorithms that allow computer programs to automatically improve through experience"*. In the case that concerns us, the "programs" that must improve over time are the statistical models that we design to either classify our observations, predict a numerical value of a response variable or simply group similar observations/discover patterns between variables.

We can find many categories of machine learning algorithms. Prediction and classification models fall into the "supervised learning" category because the models need the observations on which they are trained to be labeled (i.e., assigned to a response category or value). Descriptive models, which try to determine relationships between variables without having a target, are classified as "unsupervised learning". The k-Nearest Neighbors and the Support Vector Machines algorithms fall into the "supervised learning" category; the Random Forest model is considered to be a "Meta-Learning Algorithm" (focused on learning how to learn more effectively).

But how do machine learning models learn from our data? In a sort of way, they mimic the human process of learning (i.e., collecting data, converting it to the abstract concept that represents, generalizing the abstractions to create knowledge and finally checking if said knowledge correctly represents new observations). Each of the models has its own way to learn from the data and improve its performance, and while some of the methods show a high degree of transparency in their learning method, others are considered a "black box".

Support Vector Machines

The SVM models are classificatory algorithms that can be used for both classification and regression, but the former is the most common of their uses (binary classification).

The algorithm tries to represent the observations in a tridimensional space such that a flat boundary can be drawn separating observations from different groups. Said boundary is known as hyperplane and is designed to leave in each of its sides the maximum number of well-classified observations.

If observations can be placed in a tridimensional space and be separated by a flat surface, they are called "linearly separable". If this is the case, the model has to face a new question: which of all the possible hyperplanes will better separate future observations?

To select the best hyperplane, the algorithms searches the "Maximum Margin Hyperplane (MMH)", which is the one that generates the greatest separation

between classes. The observations that are closer to the MMH are called “Support Vectors”, and the model uses complex vector geometry mathematics to determine the MMH from them.

If only two classes are involved in the classification process, the support vectors are easy to find. A perimeter is drawn for the outer samples of every class (known as convex hull), and the shortest distance between perimeters is found. The observations that generate said distance are the support vectors and the hyperplane is calculated from them.

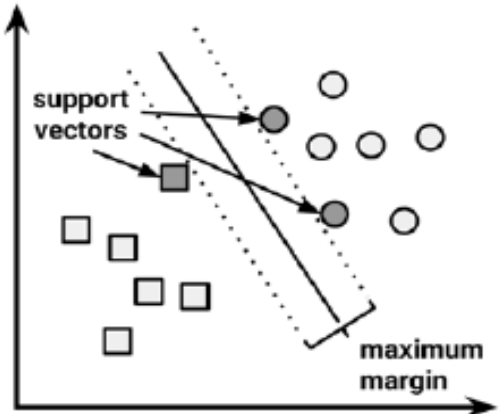


Figure 3: Linearly separable data in a SVM. In the image, squares and circles represent the two categories of the response variable; the solid line represents the MMH, and the colored observations are the Support Vectors.

If by all means there is always an observation (or more) that is misclassified by the hyperplane, then we are working with “nonlinearly separable data”. If this is the case, a hyperplane is calculated in as similar way than in the previous example, but now a penalty is assigned to all the misclassified observations (a cost parameter, “C”).

One of the keys that makes SVM so appealing is the representation of the observations into higher dimension spaces. The transformations into different dimensions are called “kernels” (or kernel tricks), and do not need to be linear. So, by trying different kernels, we may be able to draw a hyperplane to observations that were originally not linearly related.

Basically, kernels have the potential to discover mathematical relationships between variables, learning trends that were not explicitly represented by the original data.

Some of the most common kernels are the linear, the polynomial, the sigmoid and the RBF (radial basis function). Other available kernels in the R package “kernlab” are the Laplacian, Bessel, ANOVA, Spline and String.

According to literature (Lantz, 2015) the SVM models are not prone to overfit and thus perform well over new observations, and have overall good accuracy. Despite their potential, they give almost no feedback about the process that has been used to classify the observations (black box) and should not be used if transparency is an important factor in the statistical study. Also, trying different kernels to find the best performance can be time consuming.

In this study, the SVM models will be generated with the “ksvm{kernlab}” function. There are two parameters that can be used to increase the model performance, which are: **kernel** (defines the type of kernel to apply) and **C** (the cost parameter).

k-Nearest Neighbors

The k-Nearest Neighbors is a classification algorithm that assigns a category to the new observations based on their similarities with the observations on which the model has been trained. The algorithm bases its theory on the assumption that individuals of the same class must share similar traits.

In the simplest of the events, we have two explanatory variables. In that situation, observations can be visualized in a scatterplot and the distance between observations "a" and "b" can be calculated using the Pythagorean theorem (Euclidean distance).

$$dist(a,b) = \sqrt{(a_{var1} - b_{var1})^2 + (a_{var2} - b_{var2})^2}\,; \quad (formula\ 1)$$

If the number of explanatory variables is greater, we no longer can visualize distances in a graphic, but we can calculate them with the Euclidean distance, using a similar formula. For "n" variables:

$$dist(a,b) = \sqrt{(a_{var1} - b_{var1})^2 + [\dots] + (a_{var\,n} - b_{var\,n})^2}\,; \quad (formula\ 2)$$

The k-NN model calculates the distance between a given unclassified observation and all the observations in the test database. Once the distances have been calculated, the model observes the "k" (any given number, decided by the statistician) observations that are closer to the unclassified sample and retrieves their class. Finally, a class is assigned by majority voting.

Since decisions are made based on the distance between observations, we need to face a problem: variables with larger ranges are prone to dominate the model. For this reason, all variables must be scaled before classification. A common scaling transformation uses the mean and the standard deviation as follows:

$$scaled\ X = \frac{X - mean(X)}{sd(X)};\ \ (formula\ 3)$$

Another scaling approach uses minimum and maximum values instead of mean and standard deviation. This leads to a possible situation in which future observations can have values outside the range of the initial population. For this reason the approach used in "formula 3" is preferred.

Modifying the value of "k" can drastically change the classification output of the model. While larger values tend to reduce the variance caused by the noise, they also make the model more insensible to subtle (but relevant) patterns (bias-variance tradeoff).

While this is one of the simplest machine learning techniques, it is still one of the most used. The model does not make assumptions about the distribution of the variables, as it takes decisions based on the distances between observations.

Despite its simplicity, the level of transparency of the model is limited, and in most cases we will not know on what variables has the model based its decisions.

Random Forests

Before we describe random forests, we must talk about the concept of bagging, or bootstrap aggregating. This technique generates a finite number of training subsets of the original database via bootstrap, and trains a model on each of said subsets. Results of every individual model are combined into a single output, either by majority voting (classification) or by taking the mean value (prediction).

Modeling of bagging sets can be achieved via multiple algorithms, but one of the most common is by using decision trees. Decision trees classify observations by making decisions over variables, and for this reason they tend to suffer great modifications even to the smallest changes in the input data. This variability is used by the bagging approach to ensure that observation diversity is represented in the final model. The bagging model based on decision trees is called random forest.

The first step of the modeling is training the random forest. In this step the model generates “n” decision trees, each over a different subset of the training dataset. Here, “n” is the number of trees parameter, decided by the statistician: larger numbers of “n” will assure that every observation is predicted at least a few times (stabilizing the Out Of Bag error), in expenses to computational power. A second parameter, called “mtry” is also needed: this value limits the number of variables that are used in each split of the decision trees.

Once the decision trees have been trained, they are tested with all the variables that were left out in the bagging process (i.e., every variable of the training database is used to test all the decision trees that have not been trained with it) and a response is calculated for every one of them: either a category is assigned (if response variable was categoric) or a value is predicted (if it was numeric). At the end of the training process, the model returns the OOB error, which can be interpreted as a good estimate of future predictions.

In fact, if we predict new observations with the generated model we expect to obtain similar accuracy values.

Random forests are considered top-tier models, useful for both classification and regression. Said models can work with both numeric and categoric data, and perform well no matter how big the dimensionality of the dataset is. On the other hand, understanding their decision processes is not a simple task (often not even doable): decision trees are easily interpretable, but pooled trees lose this interpretability. However, we can infer variable importance in a random forest

model based on the number of decision trees that used each variable to make their decisions.

The criteria used in this study to determine the most important variables in a random forest are the mean decrease in accuracy and Gini.

The former is calculated during the OOB error calculation phase: variables importance is stablished by removing them (one at a time) from the decision tree models and observing the variance in accuracy. Variables with larger mean decrease in accuracy are more important for classification/regression.

The mean decrease in Gini is related to the contribution of each variable to the purity of the tree nodes. Every time a split is made in a decision tree, Gini coefficients are calculated for child nodes (based on the homogeneity compared to the original node), and nodes from every variable are summed and normalized. Higher values of the mean decrease in Gini represent the variables that more purely classify data.

4. Materials and methods

4.1 Management of data and statistical analysis (RStudio/Rmarkdown)

All the data processing and studying has been conducted using R and can be found in the adjacent document “DEXAmarkdown.Rmd” (or the .pdf version of it). This is a “markdown” formatted document that contains the code used during this journey, along with the minimal text indications necessary to follow and understand it. Because of the extensivity of the document (especially when exported into a PDF file) it is presented as a side product of the analysis and is not included in the main document.

Graphics, tables and other visual and/or numeric outputs can be generated with the markdown document but are also provided as an annex.

Details about the R and RStudio versions, along with a complete list of the loaded packages, can be found in the Annex 1: R packages and versions

4.2 Initial treatment of data

4.2.1 Merging databases

The original database contains 1480 observations of 82 variables. A second database contains some extra information from the observations, and the columns of interest are merged into the first database. The merging of the columns has been done by finding common values between “Database1 - ID” and “Database2 - historial”.

The merged columns contain information about the date the DEXA was performed and the date that the patient got diagnosed with HIV. Both columns have been merged into the first database, resulting in 1480 observations of 84 variables. An 85th variable (disease age) is generated by calculating the years that have passed between the day the patient got diagnosed with HIV and the day the DEXA was performed.

Columns have been renamed and arranged in the following order:

General information variables (1:7)

"ID", "gender", "gender_num", "Age", "Age_cat", "Height", "Weight"

Muscle/Lean related variables (8:31)

"RAFp", "RAFg", "RALg", "LAFp", "LAFg", "LALg", "BothAFp", "BothAFg", "BothALg", "RLFp", "RLFg", "RLLg", "LLFp", "LLFg", "LLLg", "BothLFp", "BothLFg", "BothLLg", "TFp", "TFg", "TLg", "TotalFp", "TotalFg", "TotalLg"

Bone related variables (32:61)

"L1BMD", "L1T", "L1Z", "L2BMD", "L2T", "L2Z", "L3BMD", "L3T", "L3Z", "L4BMD", "L4T", "L4Z", "L1L4BMD", "L1L4T", "L1L4Z", "L2L4BMD", "L2L4T", "L2L4Z", "NeckFBMD", "NeckFT", "NeckFZ", "WardsBMD", "WardsT", "WardsZ", "TrochBMD", "TrochT", "TrochZ", "TotalFBMD", "TotalFT", "TotalFZ"

Summary variables (62:85)

"BMI", "BMI_cat", "FMI", "FFMI", "Apendicularleanmas", "FMR", "FTrunkgFLegsg", "Indexdistributionfat", "FtrunkpFlimbsp", "FtrunkgFtotalg", "FLegsgFtotalg", "FlimbsgFtotalg", "LLegFgBMI", "LLegFpBMI", "Lipodistrophy", "Sarcopenia", "LipoSarcop", "phenotype", "minTscore", "Tscore_3cat", "TotalBMD", "HIV_date", "dexa_date", "Disease_age"

All variables are numeric except:

“ID”, “gender”, “gender_num”, “Age_cat”, “BMI_cat”, “Lipodystrophy”, “Sarcopenia”, “LipoSarcop” and “phenotype”, defined as factors; “HIV_date” and “dexa_date”, defined as POSIX time values.

4.2.2 Typos

Some values have been determined to have biologically or mathematically unexplainable values. Such issues have been addressed in the following ways:

- Values of 0 in weight-related variables have been replaced by NAs.
- Percentage values over 100 or proportion values over 1 are thought to be due a wrongly placed decimal separator and such values haves been divided by 100 and 10, respectively.
- Appendicular lean mass and Trunk to legs fat ratio have been recalculated with the following formulas:

$$Apendicularleanmas = \frac{BothALg + BothLLg}{Height^2 \cdot 1000}; \quad (formula\ 4)$$

$$FTrunkgFLegsg = \frac{TFg}{BothLFg}; \quad (formula\ 5)$$

- “Sarcopenia” has been redefined using recommended cutting points. Samples with appendicular lean mass values below 7 (men) or 6 (women) have been classified as sarcopenia-positive (1).
- Fat mass ratio values differ from the expected. All values have been recalculated with the following formula:

$$FMR = \frac{TFp}{BothLFp}; \quad (formula\ 6)$$

- Lipodystrophy variable has been recalculated using the FMR variable and the selected cut-off points.

- Liposarcopenia has been defined to take values of 1 if the observation has either lipodystrophy, sarcopenia or both.

4.2.3 Lipodystrophy and sarcopenia codification

Lipodystrophy diagnosis via DEXA is determined using the Fat Mass Ratio (trunk to legs fat ratio) variable. Some literature recommends a single FMR cutoff value of 1.26 (Beraldo et al., 2015), while others establish different cutting points for men and women (1.961 and 1.329, respectively) (Freitas et al., 2010). Since gender seems to have an important role in this study, the second approach is used. According to literature, said cutoff points are associated (in their dataset) with a sensitivity of 58% / 51% (men, women), 84% / 95% specificity, 90% / 90% predictive positive value and 45% / 66% negative predictive value.

Values above the cutting point (more fat in the trunk than in the legs) are associated with lipodystrophy.

With sarcopenia, literature establishes a common criteria to determine the cutting points to the appendicular lean mass variable as being 2 standard deviations below the population mean (by genders). However, the exact values are strongly dependent of the ethnicity of the population of study (Abdalla et al., 2020; Malmstrom, Miller, Herning, & Morley, 2013; Shafiee et al., 2018; Viana et al., 2018).

For our dataset, the cutting points for men and women are set to 7 and 6 kg/m^2, respectively, as recommended in literature (Cruz-Jentoft et al., 2019). Samples with appendicular lean mass below these values are diagnosed with low muscle mass (~ sarcopenia).

4.2.4 Duplicate IDs

Even though we worked with a database that supposedly contained the latest of the DEXAs of every patient, some of the entries of the database shared the same patient “ID”.

Such values have been identified, and only the most recent one has been kept. To know which entry was the oldest, column “Age” has been used.

After removing duplicate samples, the database size was 1475 x 85 (reduction of 0.34%).

Duplicate IDs removed: “148191”, “153111”, “153605”, “168789” and “232053”.

4.2.5 Calculation of “minTscore” and “Tscore_3cat”

The categoric classification (“Tscore_3cat”) has been made according to the WHO criteria. To calculate the minimum T-score, only measures from three regions have been considered: lumbar spine L1-L4, Femoral neck and Total hip.

Single vertebrae are excluded from study since L1L4 is a more reliable value. Wards is excluded because it usually overestimates the severity of the disease and is associated with false positives (doroudinia2015). Trochanteric values are taken into consideration: despite having lower bone density because of the presence of trabecular bone, the T scores are a reliable value.

4.2.6 Management of NAs

Dealing with missing values can be trivial or crucial based on the statistical models that must be used later. While some of them are capable of working with missing data, others require complete cases to work. In order to broaden the analytic possibilities, NAs have been evaluated and deleted when possible.

Samples containing NAs have been identified in order to discover any possible relationships between them. However, they do not seem to represent any subgroup of our data and are normally distributed for all other variables. Therefore, said observations are safely removed. Brief numeric description of the deleted samples can be found in the Annex 2: Summary tables (*Table 7: Summary of the deleted samples*).

For the purpose of this study, all observations containing NA values have been removed, ending up with a database of 1426 observations (3,6% reduction).

A less restrictive approach has been considered, allowing some variables to have NAs. The resulting database showed a reduction of 3,4%, while increasing the complexity of the study. The approach has been discarded, but its details can be found in the markdown document.

IDs of the deleted observations:

"35981"	"45806"	"90908"	"92242"
"93264"	"143630"	"145445"	"173955"
"175815"	"193544"	"206459"	"261221"
"273533"	"275572"	"283737"	"289085"
"305728"	"314142"	"341762"	"420494"
"447739"	"470836"	"473998"	"474320"
"486931"	"496757"	"500488"	"501443"
"502384"	"503754"	"506932"	"507355"
"510354"	"526215"	"537097"	"562458"
"624562"	"10000634"	"10006435"	"10448854"
"10451281"	"11139789"	"11144163"	"12726293"
"14511840"	"14667162"	"15257591"	"18113655"
"18121144"			

4.2.7 Outlier detection

Extreme values have been studied based on the Mahalanobis distance. Statistic has been calculated accounting for the gender effect. Distribution of the Mahalanobis distances have been visually represented using scatter plots, and the 5 observations with the most extreme values have been manually studied to

check for any incongruences or patterns. With the available background information, no assumptions can be made and all observations have been kept in the study.

4.3 Exploratory data analysis

An exploratory analysis is performed to determine the nature of the variables, their distribution and to identify possible correlations.

Proportions inside the levels of every factorial variable have been observed using bar plots and proportion tables. The graphics represent the proportions inside the whole group of observations and inside every gender.

Pie plots are strongly ill-advised in literature because of their ambiguous interpretation; instead, bar plots have been used.

Bar plots have been generated using the package “ggplot2”. Proportion tables have been generated using both “dplyr” and “base” packages.

Numeric variables have been firstly approached with summaries (*Annex 2: Summary tables: Table 8 and Table 9*). Two tables have been designed, containing basic numeric information about the variables (minimum and maximum values, 25%, 50% and 75% percentiles, mean and standard deviation). First table refers to the whole observed sample, while the second table is separated by gender. Min and max values have been excluded from the second table for brevity.

Tables have been generated with “dplyr” and “knitr” packages.

Gender is considered to have an important role in the database, as it is known and documented that body distribution of lean and fat tissues is strongly dependent of it (Karastergiou, Smith, Greenberg, & Fried, 2012). In order to determine the role that gender has in our dataset, all variables have been studied via statistics (t-test) and graphics (density plots).

A T-test has been applied to every numeric variable to check from mean differences due to gender. The null hypothesis was: “Ho: Differences in the sample means cannot be seen, if we separate samples by gender”. A significance level of 0.05 has been established. Variances of samples have been treated as unequal, and therefore a Welch T-test has been conducted.

The package “stats” has been used to perform the test.

Also, a density plot has been calculated for every numeric variable. Distinction between genders has been considered and included in the resulting plots: blue and red colors represent the male and female observations.

Graphics have been generated using “ggplot2”.

Linear correlation has been checked using correlation plots (code font: Williams, 2020) Factor variables have been considered as numeric. A correlation threshold of 0.5 has been established when plotting the results: only variables with values over the threshold are part of the resulting graphics.

Correlation has been studied between all variables and inside every subgroup (bone, lean/fat and summary variables).

Finally, a normality test has been conducted using the Shapiro-Wilk method. This method is known to be very restrictive, and although variables may not pass the test, it is usually safe to treat them as normally distributed.

Normality has been checked under two scenarios: for the whole samples and by genders. Significance level has been established at 0.1, as suggested in literature (Royston, 1995).

Visual inspection of variable normality has been performed with quantile-quantile plots, using the package “car”. Graphics have been generated for all samples and by genders.

Various transformations have been considered to improve normality, i.e., log, exp, sqrt, inverse, sin, cos, x^2. Other methods (such as Tukey’s Lambda or Box-Cox) have not been applied due to the difficulty of interpreting the biological meaning of the resulting transformations.

4.4 PCA

Principal components analysis has been conducted with the numeric variables of the dataset. Variables have been transformed, normalized, and standardized prior conduction of the analysis. Proportion of variance explained and cumulative proportion have been used to decide the number of PC needed to safely describe our data.

The following variables have been excluded from the PCA, in an effort to work only over the pure (i.e., non-arithmetically calculated) variables:

Lipodystrophy	BMI	Indexdistributionfat
Sarcopenia	FMI	FtrunkpFlimbsp
LipoSarcop	FFM	FtrunkgFtotalg
Phenotype	Apendicularleanmas	FlegsgFtotalg
minTscore	FMR	LLegFgBMI
Tscore_3cat	FTrunkgFLegsg	LLegFpBMI

Biplots have been generated to represent component pairs 1-2 and 3-4. Each plots’ observations have been colored by groups to identify patterns.

4.5 Graphical models

Further exploration of the correlations in our database has been managed with graphical models. A first approximation has been conducted using undirected and directed Gaussian Graphical Models (GGMs), followed by an exploratory analysis including factorial variables via Mixed Graphical Models (MGMs).

We must keep in mind that GGMs rely on the assumption of normality of the data (Bhushan et al., 2019), which is not strictly achieved in our dataset. Untransformed variables have been used for this analysis to simplify visual interpretation of the results. It's also remarkable that all assumptions made observing the GGM graphics are merely hypothetic and do not confirm causal relations (Epskamp, Waldorp, Mõttus, & Borsboom, 2018).

GGMs have been designed by genders, excluding all non-numeric variables. Colors have been assigned to represent the nature of the variables (patient information, fat/lean, vertebrae, femoral, summary, minTscore, TotalBMD and disease age). To calculate the sparse estimation of the covariance matrix, the EBIC glasso algorithm has been used, with a gamma of 0.5. Edges with weights below 0.1 have been excluded from the graphic.

Graphics have been plotted using the Fruchterman-reingold algorithm, in which each node repulses each other, but connected nodes are also attracted (thus forming clusters).

To direct the graphics, the functions "skeleton, udag2pdag{pcalg}" have been used. Fitness of the resulting DAG (Directed Acyclic Graphics) has been tested with a chi squared test.

The function "mgm{mgm}" has been used for the Mixed Graphical Models. The variable "phenotype" is not part of the analysis due to the extreme lack of observations in various of its levels. Tunning parameters have been estimated with two approaches: via EBIC glasso (gamma 0.5) and via 10-fold cross validation. A few subset techniques have been considered to design the Mixed Graphical Models, such as selection of the 10 more important variables of the first Principal Component, selecting the most relevant variable of the 10 first principal components, selecting 3 variables of every group of measures (fat, lean, bone mass, etc.). A graphic has also been conducted with all the variables of the study.

4.6 Directed analysis for osteoporosis and osteopenia

All data partitioning has been done using "createDataPartition{caret}". This function allows for an equal distribution of the response variable between groups. Size of the training split: 2/3.

Proportion between classes in the response variables has been evened (when needed) using the SMOTE over-sampling technique ("SmoteClassif{UBL}"). This algorithm applies both an over-sampling of the less represented class and an under-sampling of the most represented one.

4.6.1 Random forests

To test how well our data can predict the osteoporosis risk of the observations, random forests have been performed in duplicate: with all the variables and excluding those that directly explain the response (i.e., bone variables from L1BMD to totalFZ and totalBMD).

Random forest models have been conducted for both regression (over minTscore) and classification (over Tscore_3cat). Number of trees has been established seeking best performance (computational processing is fast and therefore not an issue). 1 to 25 variables have been considered at each split (mtry value).

Once the parameters have been optimized, the models have been used to predict over the "test" data split and the performance has been evaluated with ROC curves, AUC and RMSE.

In aims of improving performance, a classification model has been tested accounting for only two factor levels: healthy bone vs diseased (osteopenia and osteoporosis collapsed in a single category).

4.6.2 Support Vector Machines

Support Vector Machine algorithms have been implemented in an attempt to improve the model performance when classifying observations by Tscore_3cat. Function used: "ksvm{kernlab}".

Various kernels and penalty values (C) have been implemented, seeking the best accuracy. Kernels tested: linear, polynomial, Gaussian, ANOVA, hyperbolic and Laplacian; penalty values ranged from 1 to 10. The analysis has been conducted over transformed and standardized data. Two models have been tested, classifying into 3 and 2 categories (healthy/osteopenia/osteoporosis; healthy/diseased).

4.6.3 k-NN

For the k-NN algorithm the number of neighbors has been set seeking the best accuracy. Range of values tested: 1 to 10 neighbors. Only numeric variables have been considered in the algorithm, and observations have been classified into two categories (healthy and diseased). Analysis has been conducted twice, balancing the categorical variable and without doing so. All numeric data has been centered and scaled prior classification, to ensure equal importance of variables (independently of their range). Package used: "knn{class}".

5. Results

5.1 Univariate study

5.1.1 Factorial variable bar plots

Plots can be seen in Figure 4. We observe a higher percentage of patients below 50 years old for both women and men. Despite being a young database, more than the 70% of the women suffer from either lipodystrophy or some sort of muscular mass malfunction.

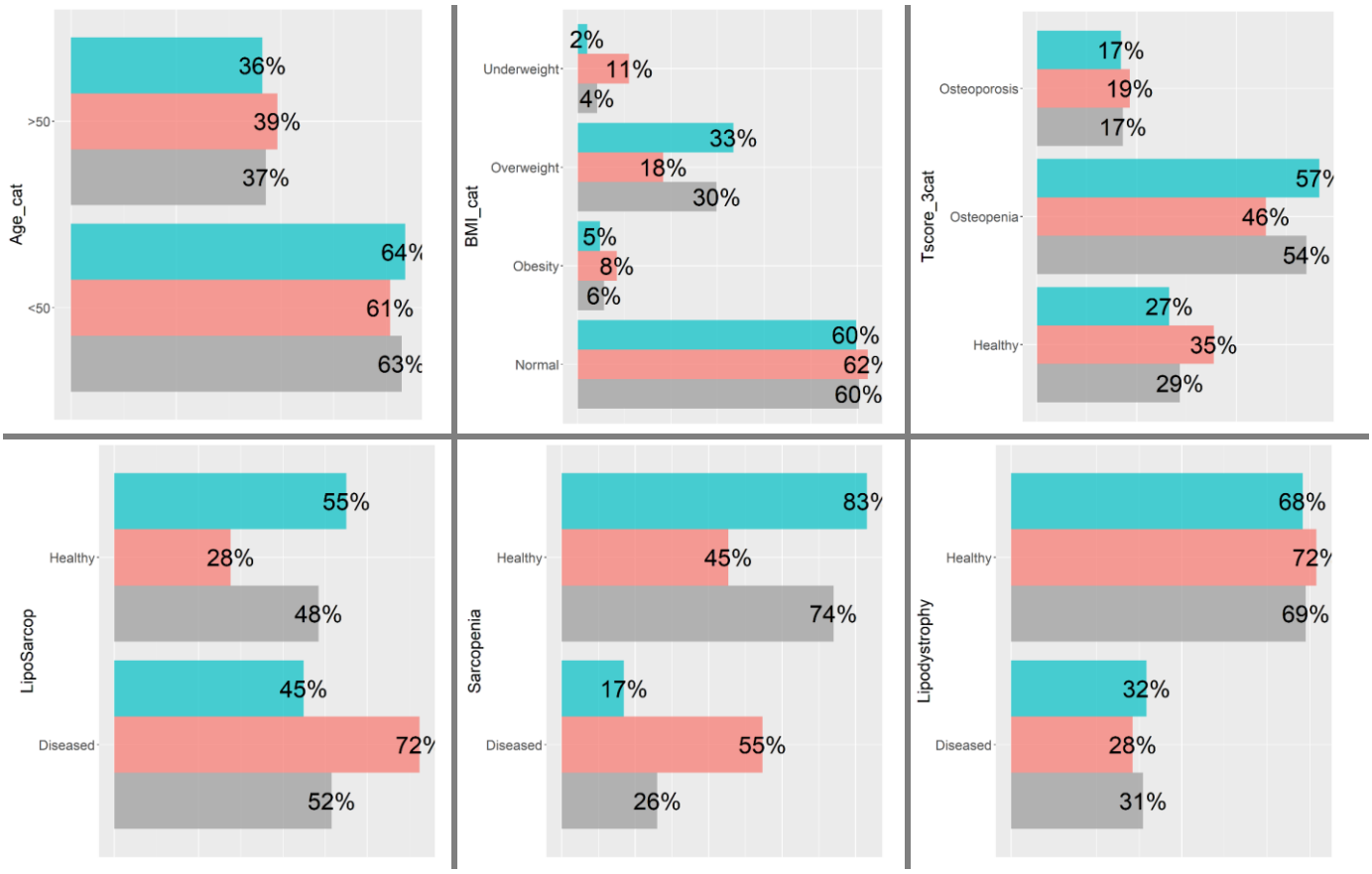


Figure 4: Bar plots for factorial variables. From left to right and top to bottom: Age (above 50 or under 50), BMI (underweight, normal, overweight or obesity), T-score classification (healthy, osteopenia or osteoporosis), presence of a) lipodystrophy or sarcopenia, b) sarcopenia and c) lipodystrophy (1: presence, 2: absence). Colour code: blue-men, red-women, gray-all samples.

Low mass muscle is only observed in 1 of every 5 male patients, while half of the female observations suffer from it. Lipodystrophy is equally present in both genders, at about a 30%.

10% of women present both lipodystrophy and LMM, while a 28% do not show any of the comorbidities. For men, those values are of 4% and 55%.

Examining the categorical T-score we can infer that women tend to have more extreme values, i.e., they are more represented in the healthy and osteoporotic groups. In the other hand, men are more likely to present moderate osteopenia.

BMI distribution is uneven, and the underweight and obesity groups are strongly underrepresented.

5.1.2 Summary overview

The summary tables can be found in Annex 2: Summary tables (*Table 8: Summary of the variables* and *Table 9: Summary of the variables, by gender*). We are studying a middle-aged population formed mainly by men (~76% of the observations). The most prevalent comorbidity is osteopenia/osteoporosis (>60%), followed by lipodystrophy (30%).

Sarcopenia seems to be more present in females, while the other diseases are equally distributed.

5.1.3 Gender effect

Most of the variables of the study show a significant difference in both mean and distribution based on gender. Such effect can be seen in the following t-test results, density plots and in the Principal Components analysis (5.1.6 PCA). This observation is supported by biological facts. Therefore, in subsequent analysis will be conducted separately by genders.

T-tests

No differences between genders has been found for the T-score variables (T-test p-values >0.05). All other variables show different means between men and women. Details of the p-values can be seen in Table 2:

TotalFT	0.9803	**NeckFT**	0.2752	**Age**	0.0332
L1L4T	0.9794	**WardsT**	0.2429	**L3BMD**	0.0177
L2L4T	0.8857	**L3T**	0.23	**WardsZ**	0.0063
L4T	0.6965	**minTscore**	0.1369	**TFg**	0.0059
NeckFZ	0.455	**TotalFZ**	0.1049	**L4BMD**	0.0023
L2T	0.3006	**L1T**	0.0439	Rest of var.	< 0.0001

Table 2: T-score p-values. P-value results for every variable. Significance level: 0.05 (values below are colored in red). Variables with p-values smaller than 0.0001 are excluded from the table.

Density plots

Bone-related variables are less determined by genders. BMI and T-scores show almost no mean differences (despite the t-test results) while Z-scores are mildly influenced by gender (women tending to have higher means than men). In Figure

5 “c” the L1L4 T-score values are represented, showing no variability between genders.

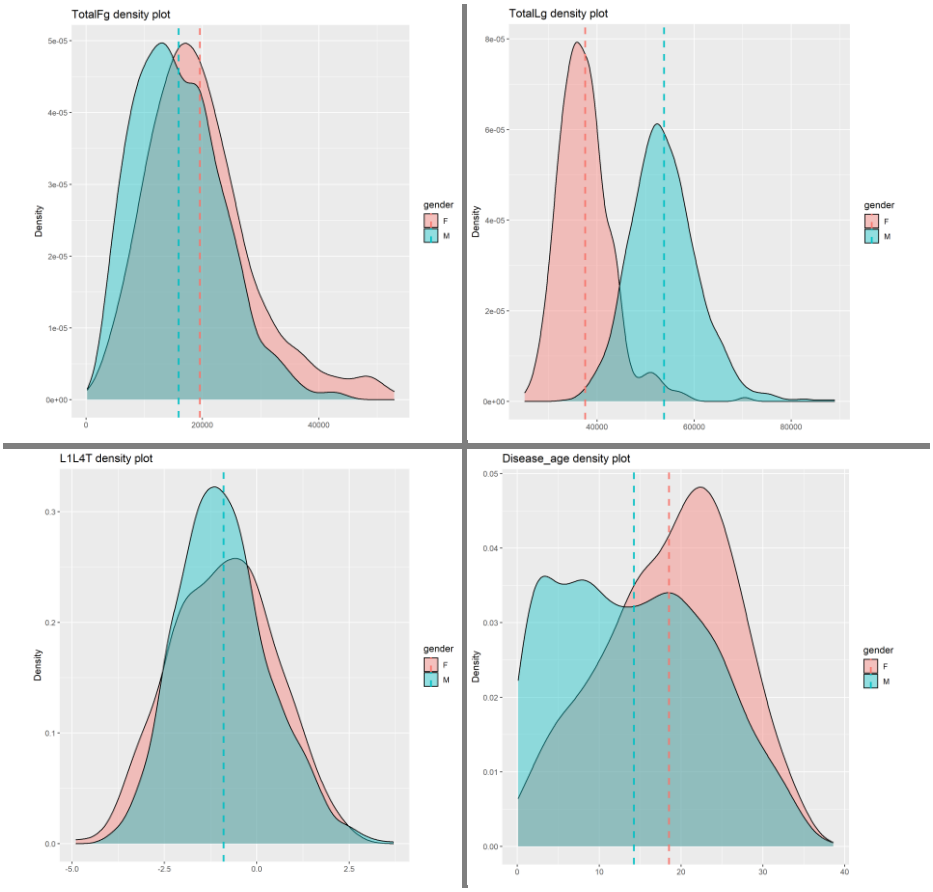


Figure 5: Density plots. a) Total body fat, in grams; b) total body lean, in grams; c) L1L4 T-score and d) Time passed from the disease diagnosis to the DEXA, in years. Blue and red colors represent men and women observations; dotted lines mark the group means.

Differences due to gender are especially significant for fat and lean related variables. Percentages show more differences than weight measurements. Differences between genders are consistent for all body parts (symmetry is preserved). While women display higher fat values for all measurements, men have higher values for all lean variables (see Figure 5 “a” and “b”).

Men tend to be taller and more robust. Age (despite the T-test p-value of 0.03) seems to be evenly divided among genders.

Women of this study have been living with the disease for longer than men. The image not only shows a difference in means, but a right-skewness for women and a left-skewness for men (Figure 5 “d”).

All the density plots can be observed in detail in Annex 3: Density plots.

5.1.4 Linear correlation

Correlation within variables is a common issue in DEXA analysis (human bodies tend to follow high levels of symmetry and proportionality and therefore muscle, fat and bone values are strongly correlated).

Correlation between bone-related variables is strong. Two groups can be appreciated (Figure 6, left): vertebrae and femoral variables. Inside each group, BMD, T and Z-scores show lineal correlation values over 0.75. Both groups are correlated to the TotalBMD variable.

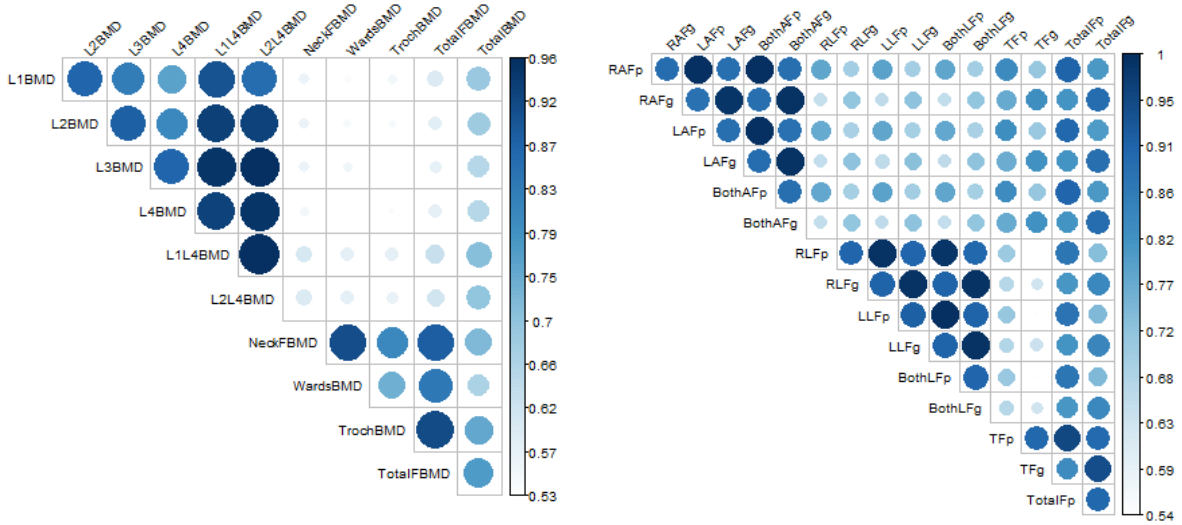


Figure 6: Bone and fat variables correlations. Left: bone variables. Only BMD measurements plotted for clarity (T-scores and Z-scores share the same behavior). Upper-left cluster represents vertebral variables (L1BMD to L2L4BMD), while the bottom cluster refers to the thigh bone (NeckFBMD to TotalFBMD). **Right:** fat variables. A first correlation cluster is observed within upper-limbs variables (RAFg to BothAFg), and a second cluster among leg variables (RLFp to BothLFg). Whole body variables (TotalFp and TotalFg) are correlated to all other fat variables.

Fat related variables are also strongly correlated (Figure 6, right). We observe higher correlation inside the arms variables, and again inside the leg-related variables. Also, a strong correlation (values over 0.9) is appreciated between left leg/arm values and their corresponding right leg/arm values (body symmetry). Total fat variables are correlated with both upper and lower limb groups and might be good options to summarize the information contained in the other variables.

Lean variables are strongly intercorrelated but are linearly independent from fat variables (Figure 7, left).

Among the “summary” variables, FMI, FFMI and Appendicular lean mass are the least correlated variables. A strong linear dependence (either positive or negative; always over 0.9) can be observed between the rest of the variables (Figure 7, right).

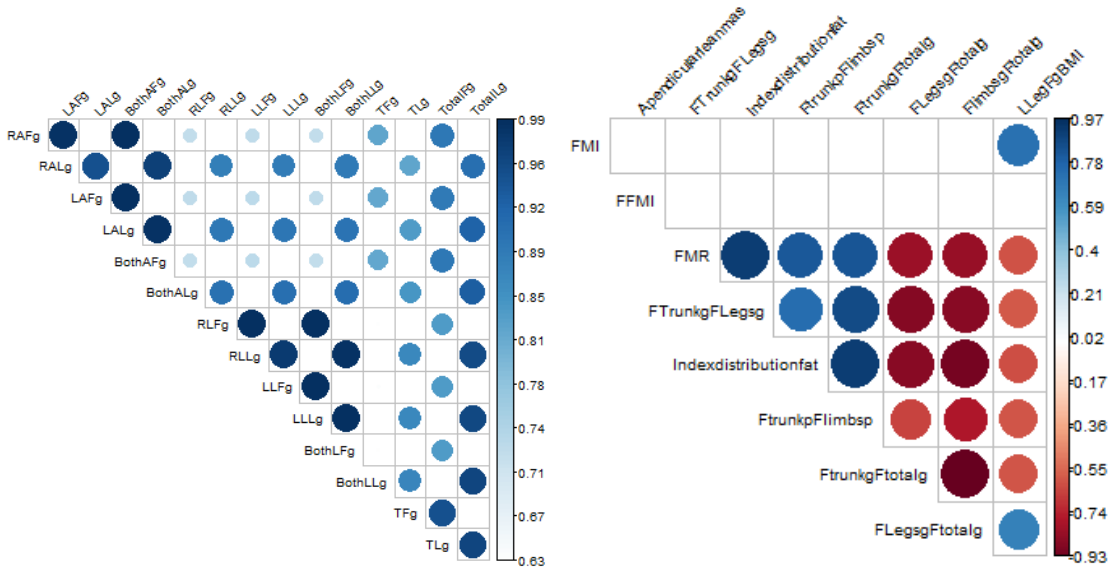


Figure 7: Lean and summary variables correlations. Left: lean and fat variables. Strong correlations can be observed between lean variables and other muscular values, but no correlations are appreciated with fat variables. **Right:** summary variables. The levels of correlation are more polarized.

5.1.5 Normality

According to the Shapiro-Wilk test, no variables present normality for neither men nor the whole population. 30 variables are normally distributed in the female subgroup:

RAFp	LAFp	BothAFp	RLFp
LLFp	BothLFp	TotalFp	L1BMD
L1T	L1Z	L2BMD	L2T
L2Z	L3BMD	L3T	L3Z
L4Z	L1L4BMD	L1L4T	L1L4Z
L2L4BMD	L2L4T	L2L4Z	NeckFBMD
NeckFT	TotalFBMD	TotalFT	FlimbsgFtotalg
minTscore	TotalBMD		

Quantile-quantile plots give us a better interpretation of the distributions and the transformations that might improve their normality. Plots are not shown in this document but are available under request.

Fat and lean related variables measured in grams have been transformed using a logarithmic scale. Exceptions: TFg, TLg and TotalFg, transformed by their root square.

BMI, FMI, FMR, FTrunkgFLegsg and Indexdistributionfat have been transformed using the logarithmic transformation. FFMI and FtrunkpFlimbsp have been inverted (1/x).

Percentage variables did not show any normality improvements with any transformation. Bone variables do not improve with any transformation.

5.1.6 PCA

The PCA conducted over the whole dataset requires 4 principal components to explain 85% of the variability, and 10 PC to explain over the 95%. The first two principal components account for the 62% of the variability.

The first principal component contains the information related to bone density: all and only the bone variables have loadings with values over |0.1| (Table 3). This subset of variables seems to be related to the presence of low muscle mass (Figure 8, left) and the presence of bone diseases (Figure 8, right).

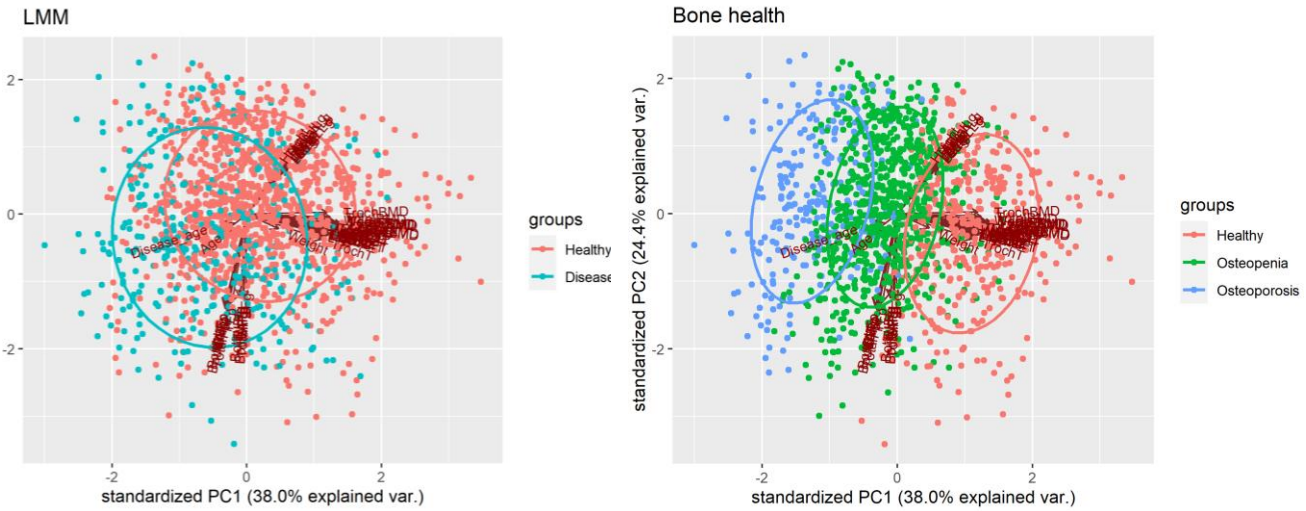


Figure 8: PCA biplots for LMM and Bone health. Colored representation of the first two principal components, using various of the categorical variables in the dataset. The first component (x axis) seems to be related to Low muscle mass (sarcopenia) and Bone health (Tscore_3cat).

The second principal component summarizes the variability of fat and lean variables, including height (as before, only fat/lean variables have loadings greater than |0.1|). This second subset of variables seems to be related to the gender effect (as we have previously seen in the T-test and the density plots) (Figure 9, left).

Subsequent principal components are biologically harder to explain. However, variables of the third principal component partially explain the BMI variable (Figure 9, right).

PCA over the feminine population reported similar results. 5 and 10 PC needed to explain 85% and 95% of the variance. First and second components divided variables in bone and muscle/fat. Second component seems to be a good explanatory variable for BMI.

PCA over the masculine population needed one more component to explain the 95% of the variability. The rest of the results are identical.

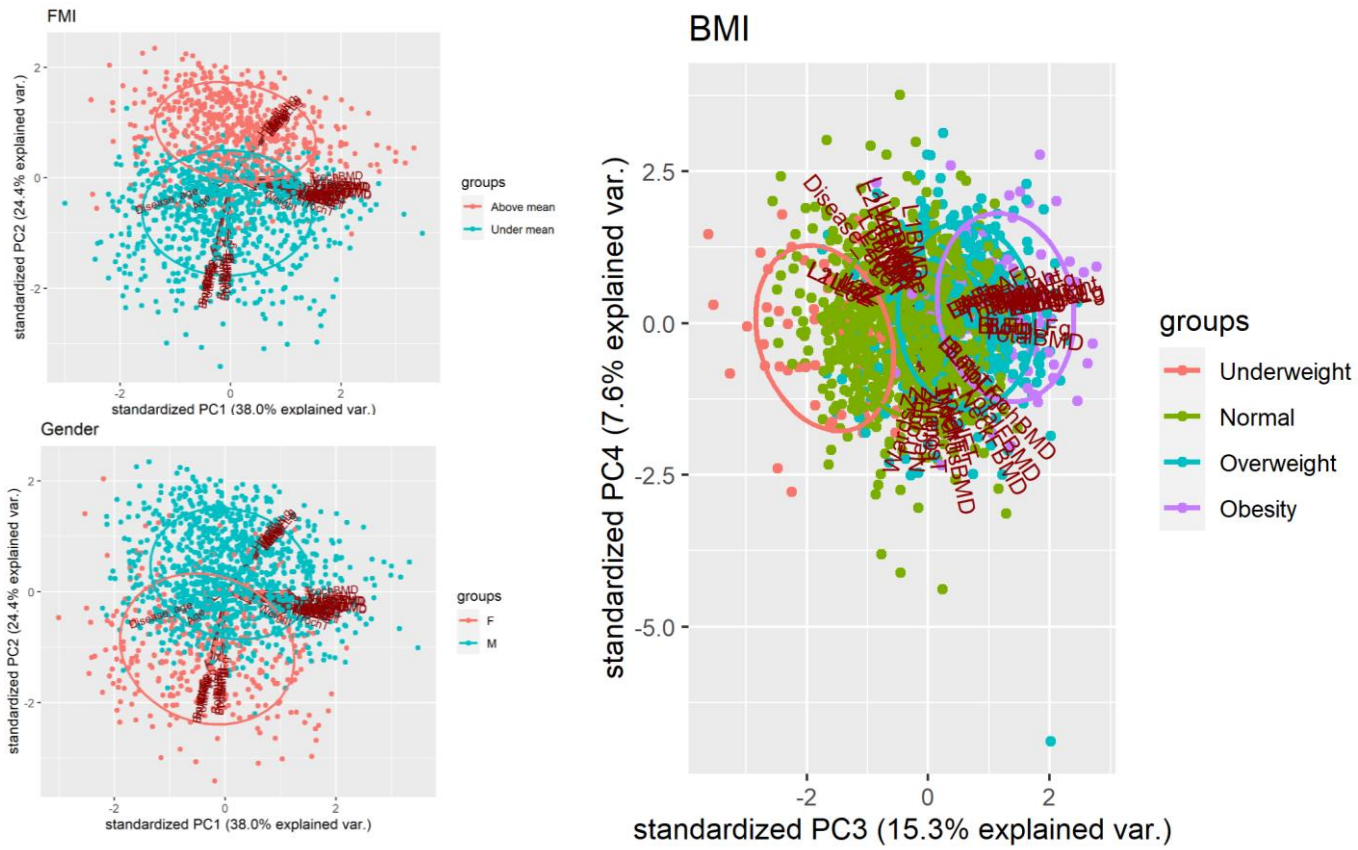


Figure 9: PCA biplots for FMI, Gender and BMI. First two PC plots are shown on the left, colored by FMI and gender. On the right, PC 3 and 4, colored by BMI.

In Figure 10 we see how the biplots for the two principal components are almost identical for men and women (vertical axis is flipped, but the interpretation does not change). The percentage of variance explained by every component is also similar.

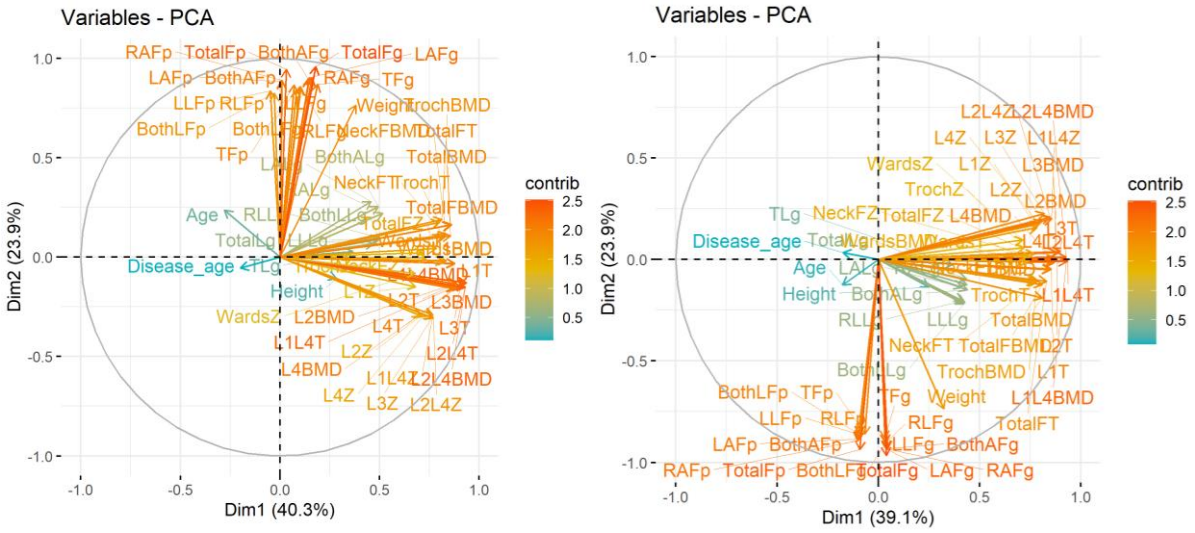


Figure 10: PC biplots, by genders. Left: components 1 and 2 of the female PCA. **Right:** components 1 and 2 of the male PCA. Variables are colored by importance.

PCA has given us new variables that contain roughly the same variability than the original database. However, because of the difficulty to biologically explain the new variables, further analysis will be conducted over the original database. No dimensionality reduction is applied (even though PCA proves it is possible).

	Pr. Comp. 1			Pr. Comp. 2		
	PCA	Fem	Male	PCA	Fem	Male
Age	0.046	0.057	0.037	0.045	0.063	0.033
Height	0.062	0.058	0.052	0.113	0.029	0.036
Weight	0.065	0.078	0.067	0.037	**0.204**	**0.196**
RAFp	0.044	0.002	0.021	**0.241**	**0.237**	**0.238**
RAFg	0.008	0.033	0.009	**0.215**	**0.241**	**0.249**
RALg	0.084	0.106	0.091	**0.127**	0.059	0.033
LAFp	0.043	0.002	0.02	**0.24**	**0.235**	**0.236**
LAFg	0.009	0.031	0.009	**0.215**	**0.241**	**0.249**
LALg	0.08	0.094	0.086	**0.123**	0.075	0.036
BothAFp	0.043	0.002	0.021	**0.241**	**0.236**	**0.237**
BothAFg	0.009	0.03	0.008	**0.214**	**0.238**	**0.248**
BothALg	0.082	0.101	0.092	**0.127**	0.068	0.04
RLFp	0.044	0.01	0.018	**0.234**	**0.222**	**0.226**
RLFg	0.014	0.022	0.007	**0.214**	**0.228**	**0.24**
RLLg	0.084	**0.102**	0.087	**0.121**	0.026	0.054
LLFp	0.045	0.01	0.02	**0.235**	**0.223**	**0.23**
LLFg	0.014	0.021	0.008	**0.214**	**0.229**	**0.24**
LLLg	0.084	0.103	0.088	**0.121**	0.025	0.058
BothLFp	0.044	0.006	0.02	**0.235**	**0.22**	**0.229**
BothLFg	0.015	0.017	0.007	**0.213**	**0.225**	**0.241**
BothLLg	0.084	**0.102**	0.089	**0.122**	0.026	0.056
TFp	0.03	0.015	0.014	**0.225**	**0.231**	**0.229**
TFg	0.002	0.039	0.007	**0.187**	**0.232**	**0.235**
TLg	0.077	0.075	0.077	**0.136**	0.007	0.006
TotalFp	0.04	0.007	0.019	**0.249**	**0.252**	**0.249**
TotalFg	0.007	0.037	0.008	**0.218**	**0.255**	**0.257**
TotalLg	0.084	0.096	0.091	**0.138**	0.018	0.026
L1BMD	**0.181**	**0.18**	**0.177**	0.032	0.008	0.013
L1T	**0.18**	**0.18**	**0.178**	0.044	0.009	0.012
L1T	**0.18**	**0.18**	**0.178**	0.044	0.009	0.012

	Pr. Comp. 1			Pr. Comp. 2		
	PCA	Fem	Male	PCA	Fem	Male
L2BMD	**0.186**	**0.181**	**0.182**	0.023	0.035	0
L2T	**0.186**	**0.182**	**0.184**	0.04	0.035	0.001
L2Z	**0.16**	**0.145**	**0.168**	0.043	0.078	0.05
L3BMD	**0.188**	**0.187**	**0.184**	0.028	0.04	0.009
L3T	**0.188**	**0.186**	**0.188**	0.044	0.04	0.011
L3Z	**0.161**	**0.157**	**0.171**	0.046	0.081	0.06
L4BMD	**0.185**	**0.18**	**0.181**	0.028	0.042	0.003
L4T	**0.182**	**0.18**	**0.181**	0.042	0.041	0.002
L4Z	**0.156**	**0.147**	**0.165**	0.045	0.08	0.047
L1L4BMD	**0.197**	**0.192**	**0.194**	0.03	0.034	0.003
L1L4T	**0.195**	**0.19**	**0.195**	0.047	0.034	0.002
L1L4Z	**0.168**	**0.158**	**0.179**	0.05	0.077	0.053
L2L4BMD	**0.197**	**0.191**	**0.193**	0.028	0.04	0.003
L2L4T	**0.194**	**0.191**	**0.193**	0.044	0.04	0.003
L2L4Z	**0.167**	**0.158**	**0.177**	0.047	0.081	0.056
NeckFBMD	**0.168**	**0.171**	**0.164**	0.012	0.03	0.028
NeckFT	**0.163**	**0.17**	**0.165**	0.05	0.027	0.028
NeckFZ	**0.154**	**0.151**	**0.155**	0.038	0.001	0.006
WardsBMD	**0.162**	**0.171**	**0.155**	0.012	0.004	0.002
WardsT	**0.158**	**0.168**	**0.155**	0.031	0.005	0.001
WardsZ	**0.146**	**0.14**	**0.149**	0.033	0.022	0.026
TrochBMD	**0.164**	**0.167**	**0.163**	0.005	0.05	0.03
TrochT	**0.154**	**0.164**	**0.161**	0.064	0.047	0.029
TrochZ	**0.144**	**0.147**	**0.15**	0.044	0.006	0.015
TotalFBMD	**0.172**	**0.177**	**0.169**	0.014	0.044	0.033
TotalFT	**0.171**	**0.176**	**0.173**	0.052	0.043	0.028
TotalFZ	**0.163**	**0.161**	**0.16**	0.03	0.012	0.007
TotalBMD	**0.171**	**0.176**	**0.169**	0.006	0.028	0.05
Disease_age	0.044	0.041	0.036	0.019	0.015	0.009

Table 3: PC 1 and 2 loadings, absolute values. PCA loadings given by the “prcomp {stats}” function, under “rotation” values. In green: variables with values over |0.1|.

5.2 Graphical models

5.2.1 Gaussian Graphical Models

In the undirected gaussian graphical models we can appreciate various patterns:

Fat and lean variables seem to form clusters of two (24-27), three (4-7-10; 5-8-11; 6-9-12; 13-16-19; 14-17-20; 15-18-21) and four (22-23-25-26) variables. Those groups of 3 variables are related to the body symmetry (left/right/both body limbs) in arms (first 3 clusters) and legs (last 3 clusters). The cluster of four variables corresponds to the total body and trunk measurements of fat, in grams and percentages. The cluster of two variables refers to the lean mass values of the whole body and the trunk.

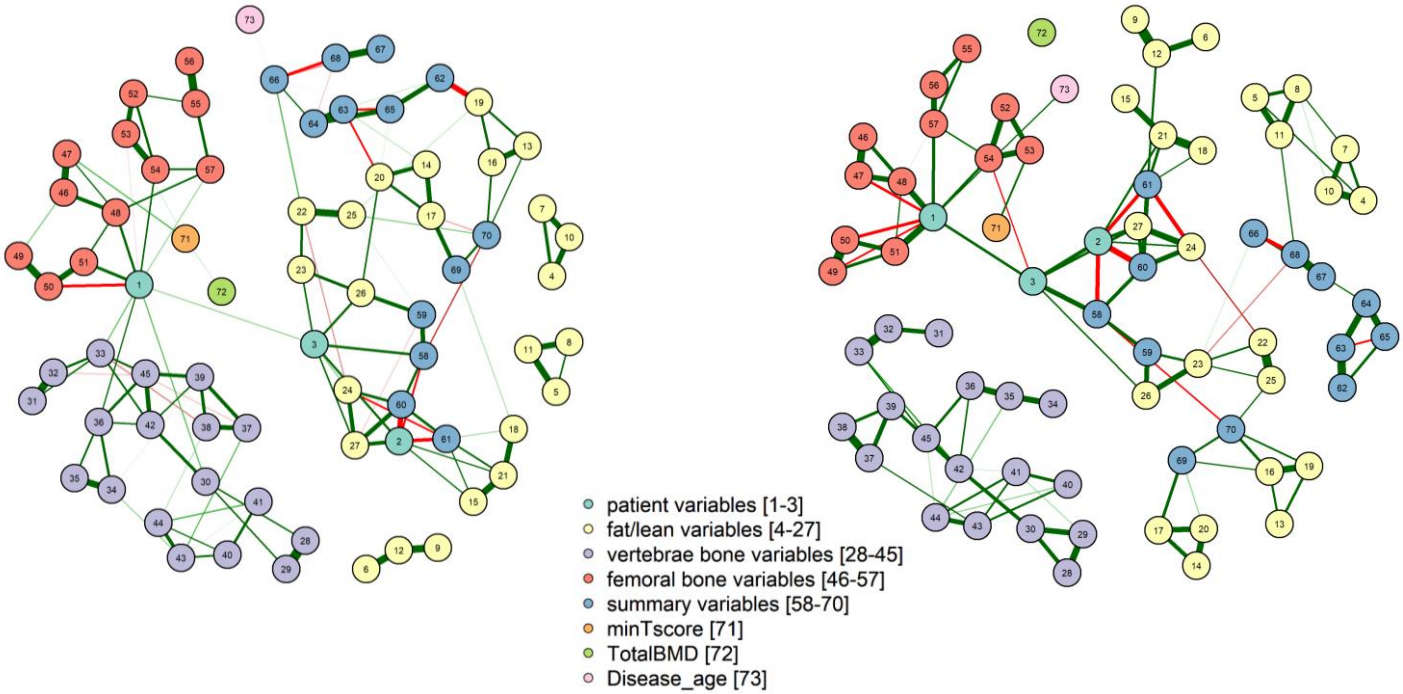


Table 4: Undirected Gaussian Graphical Models. Women (left) and men (right) variables and their correlations, as seen by an undirected Gaussian graphical model.

Vertebrae variables seem to be very intercorrelated but do not present connections outside their group. In females we can appreciate a weak relationship with the age variable (1). The edges between BMD values and their corresponding T-score are stronger than the edges with the Z-scores.

Femoral variables also form clusters of three variables, corresponding to the BMD, T and Z-scores (46-47-48; 49-50-51; etc.). Intercorrelations between groups exist, along with an external correlation to age (1) and minTscore (71). This relationship is between minTscore and NeckFT in women, and with TrochT in men. A negative edge between age and WardsT (50) seems to be consistent in both genders.

Summary variables must be studied individually, but a similar behavior is appreciated in both genders: variables 62 to 68 form a cluster, variables 69 and 70 form a second cluster and variables 58 to 61 form a more diffuse cluster. Summary and fat/lean variables are strongly intercommunicated.

There is a strong positive bound between FlegsgFtotalg (67) and FlimbsgFtotalg (68), while a strong negative bound can be observed between the last one and FtrunkgFtotalg (66). This suggests a proportionality in fat percentage between upper and lower limbs that exists independently of gender.

FMR (62), FtrunkFLegs (63), Indexdistr. (64) and FtrunkpFlimbsp (65) have a strong positive relationship. A negative edge can be appreciated in women between FMR and BothLFp (19).

Height (2) has a negative relationship with BMI (58), FFMI (60) and appendicular lean mass (61) (variables calculated as value/height^2). A positive edge can be observed between height and TotalLg (27). There is also a positive correlation between height and weight (3). Weight is also correlated to the TotalFg (26) and BMI (58).

TotalBMD (72) and disease age (73) show very weak relationships.

No relationships are appreciated between the minTscore (71) and fat/lean/summary variables.

Directed Gaussian Graphical Models show similar results, without highlighting any edges between different groups of variables, and can be seen in the Annex 4: Directed Gaussian Graphical Models.

5.2.2 Mixed Graphical Models

Mixed Graphical Models over subsets of the variables have not reported any relationships hitherto unknown. Full dataset analysis reported similar results to the Gaussian Graphical Models. When representing the whole dataset variables (Figure 11), CV reports more consistent graphics than EBIC.

In both genders, factorial variables representing presence of lipodystrophy and sarcopenia show a strong relationship via “LipoSarcop” variable, but no relationships can be observed with the rest of the dataset variables. In a similar way, categoric classification of bone healthiness is only related to the minTscore variable.

Relationships within groups can be appreciated in both genders, but no edges can be observed between bone-related and fat/lean-related variables.

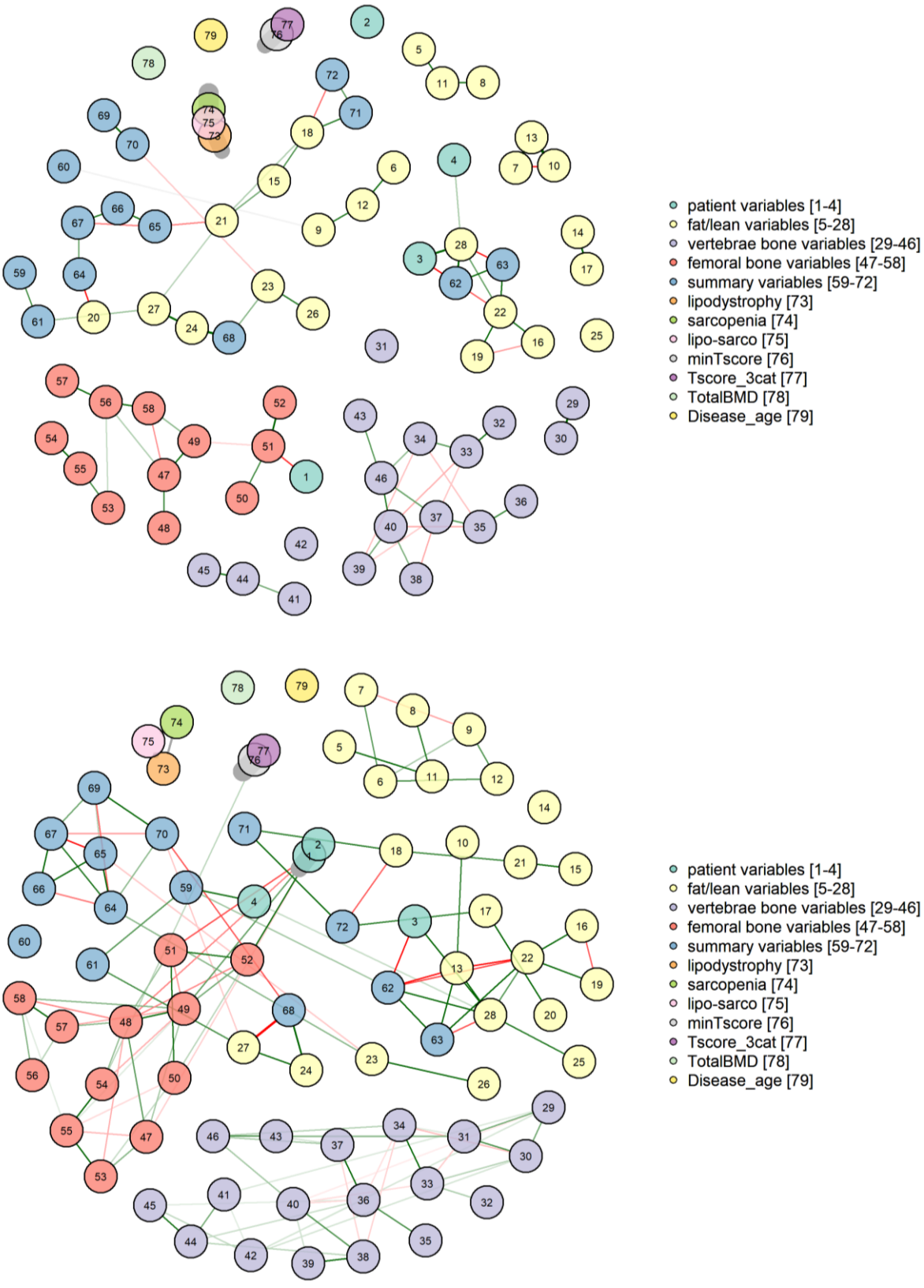


Figure 11: Mixed Graphical Models over all variables. Mixed Graphical Models for **women** (upper) and **men** (lower). Tunning parameter selected via 10-fold cross validation; nodes plotted in “spring” layout. Edge colors represent a positive (green) or negative (red) relationship; gray edges represent relationships with a factorial variable.

5.3 Directed analysis for osteoporosis and osteopenia

5.3.1 Random forest

Regression

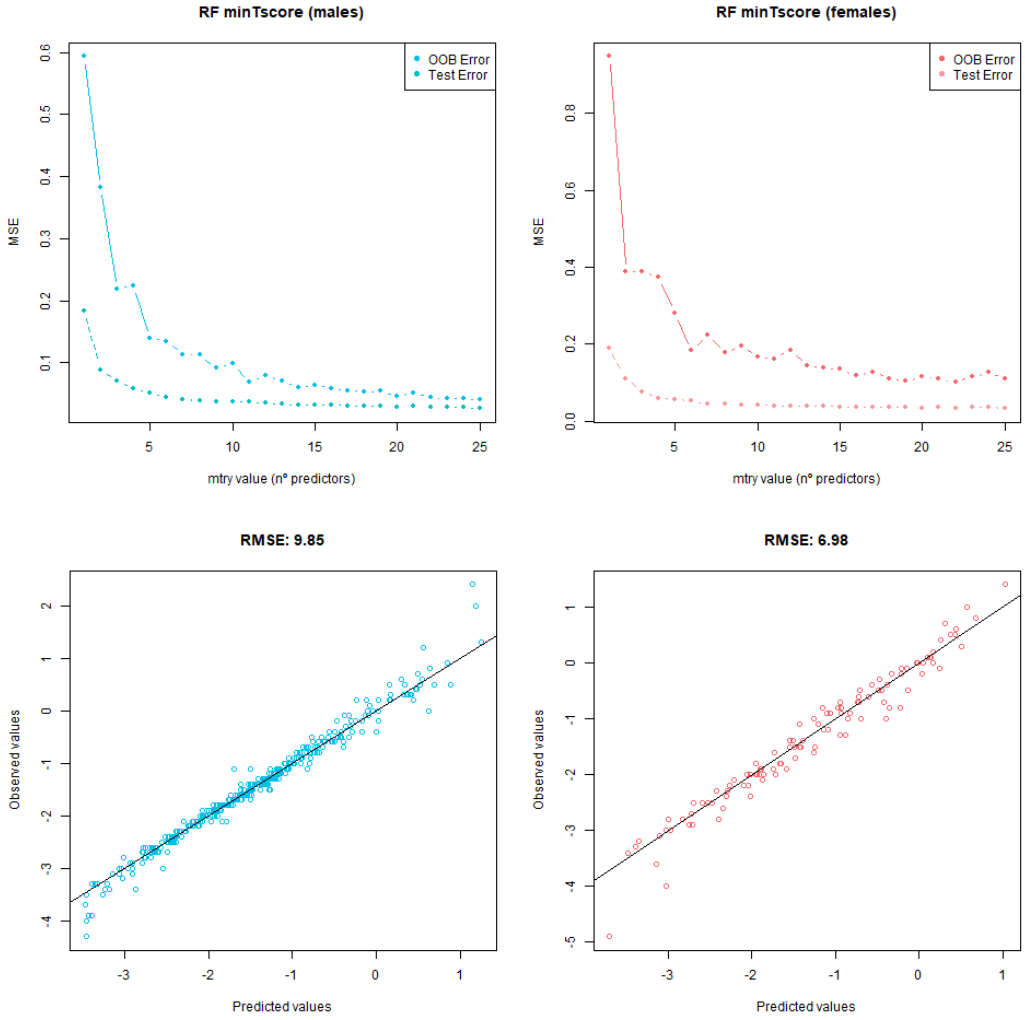


Figure 12: RF over osteoporosis, regression with all variables. Upper graphics represent the out-of-the-bag and the test errors; lower graphics represent the observed versus predicted. Men on the left, women on the right.

Forest population has been set at 100 trees (OOB error stabilized). The best number of variables at each split has been established at 11 and 13 (men and women, respectively). Adding more variables did not seem to improve the performance (Figure 12, upper). The predictions of the model are accurate, having low values of RMSE (Figure 12, lower).

When excluding the bone variables from the model, the performance decreased exponentially. Best models were obtained when using 16 and 20 variables at each split (men and women, respectively) (Figure 13). The dispersion of the graphics indicates a poor explanatory capacity of the models.

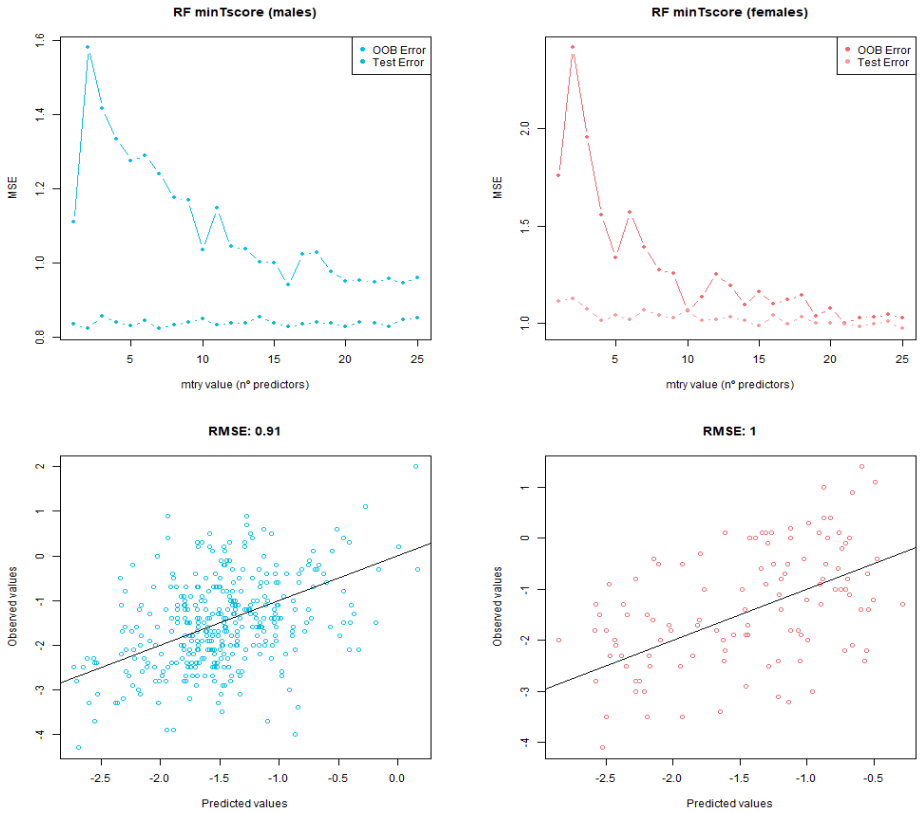


Figure 13: RF over osteoporosis, regression without bone variables. Upper graphics represent the out-of-the-bag and the test errors; lower graphics represent the observed versus predicted. Men on the left, women on the right.

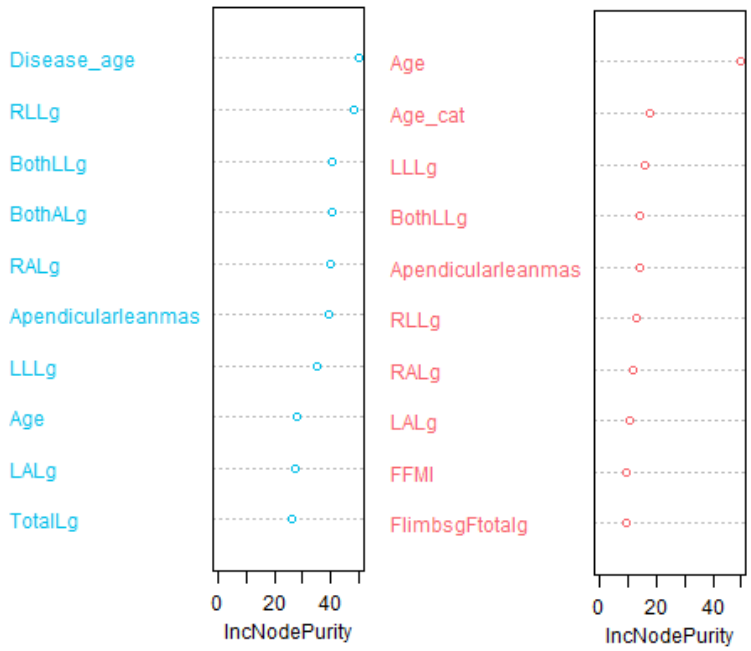


Figure 14: Variable importance in the RF. Variable importance in the regression Random forests, based on the Gini index, for men (blue, left) and women (red, right).

The variables (Figure 14) that seem to be important in both genders are related to the age and the quantity of lean mass in legs and arms. The variables that are exclusive to each of the two genders are FFMI and FlimbsgFtotalg (women) and disease age in men (along with other lean-mass related variables).

Classification

The number of trees in every forest has been set at 600. Observing the performances of the models with all the variables we observed a great classification task (Figure 15, upper). AUC values for osteopenia are 0.99 and 0.97 (males and females), and the AUC values for osteoporosis are of 0.99 and 0.98. The confusion matrix shows that most of the observations that are not correctly classified are from patients that are healthy, but the model has classified them as suffering from osteopenia.

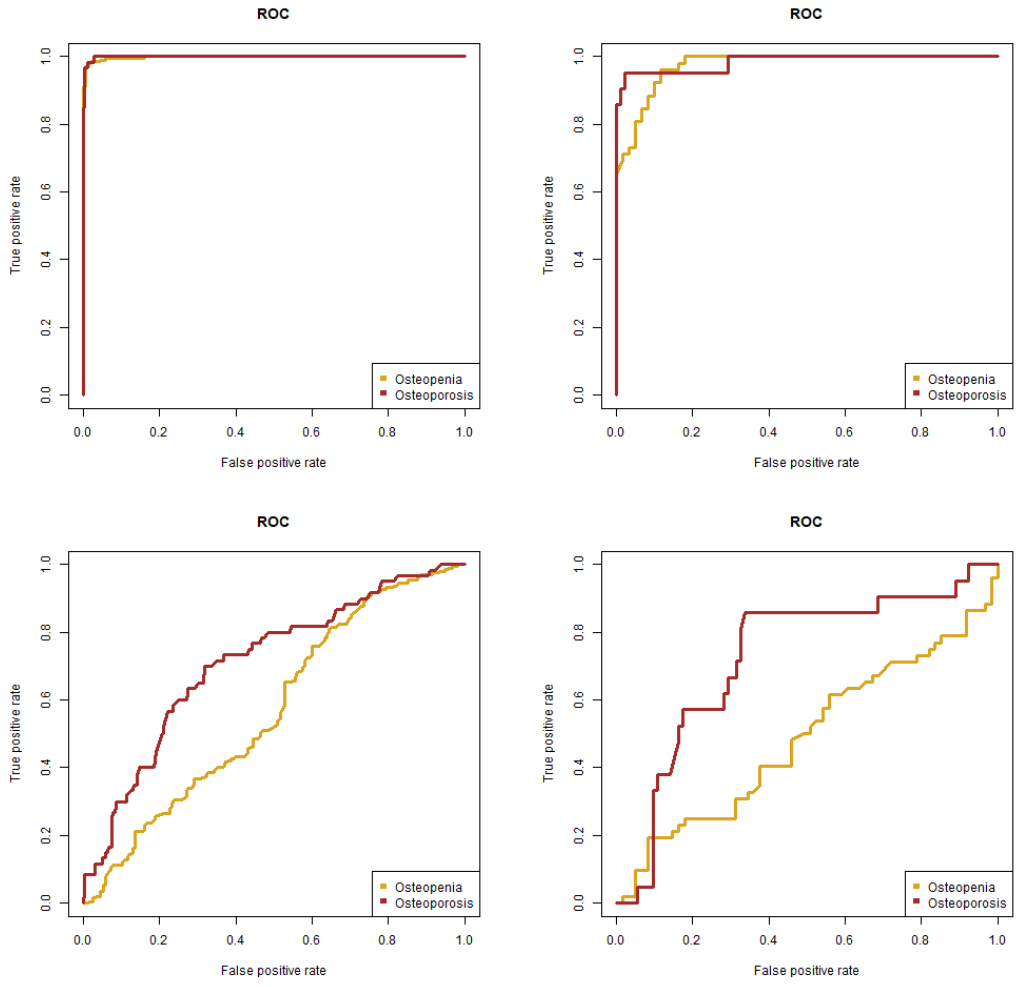


Figure 15: RF over osteoporosis, classification. From left to right, top to bottom: a) classification over males, whole variables; b) classification over females, whole variables; c) classification over males, no bone variables; and d) classification over females, no bone variables.

On the other hand, observing the results of the models without bone variables, we can see a great loss in the classificatory capabilities (Figure 15, lower), especially over the osteopenia group. The AUC values of the models are: 0.57

and 0.49 for osteopenia; 0.71 and 0.72 for osteoporosis (males and females, respectively).

The model with just two categories (healthy vs diseased bone, variables balanced) does not show a real improvement in the classification power. Although the model sensitivity is remarkably high (72% and 70%, men and women values), the specificity is lower (48% and 67%, respectively). Over half of the healthy individuals have been incorrectly classified as diseased. The AUC for the models are 0.66 and 0.73 (men, women) (Figure 16).

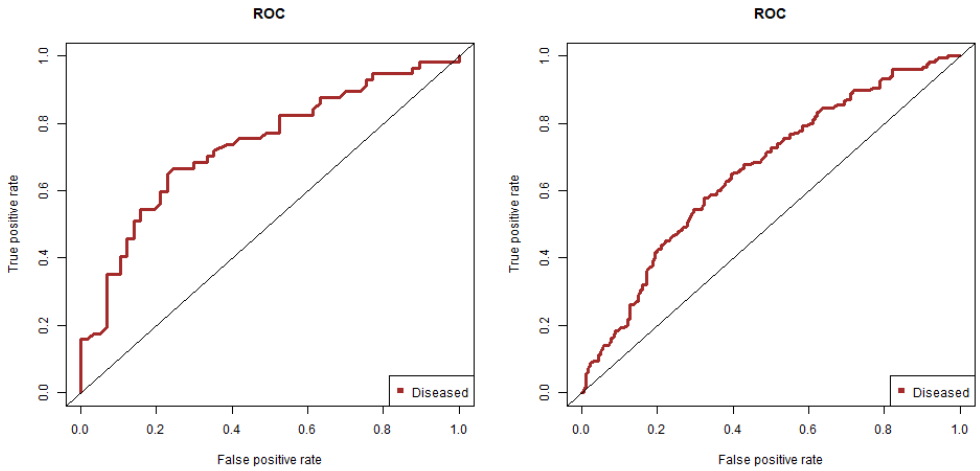


Figure 16: RF over osteoporosis (2 levels), classification. Male (left), and female (right) models of classification, based on two categories: healthy or diseased bones.

The important variables in the classification random forests (using 2 and 3 classificatory categories) are really close to the ones obtained in the regression models (Figure 14).

5.3.2 Support Vector Machines

Models over transformed and balanced data performed significantly better than their untransformed relatives. Accuracy when trying to predict over 3 categories did not report any relevant results. Classification in two categories performed relatively well in both genders (detailed values can be found in Table 5). The best kernel function had been the “vaniladot” (linear kernel) in both genders, and the best penalty values for observations falling on the wrong group had been found to be “C = 10” and “C = 4” (men and women, respectively).

	Male model	Female model
Accuracy C.I. 95%	0.72 – 0.80	0.64 – 0.81
Kappa	0.53	0.46
Sensitivity	0.78	0.75
Specificity	0.76	0.70

Table 5: SVM classification algorithms performances.

Predictive values are slightly better in men than in women, possibly because of the bigger sample size. However, the overall performances of the models are considerably good.

5.3.3 k-NN

Best performances (in terms of accuracy, kappa, sensitivity and specificity) were observed with the balanced databases. The best number of neighbors had been established at 15 for both men and women. Best model performances can be seen in Table 6.

	Male model	Female model
Accuracy C.I. 95%	0.79 – 0.87	0.71 – 0.87
Kappa	0.67	0.6
Sensitivity	0.88	0.79
Specificity	0.79	0.8

Table 6: k-NN classification algorithms performances.

Male and female models seem to perform better than the previous algorithms in classifying both the positive and the negative classes. According to the models, over 80% of the patients suffering some degree of bone disease will be positively identified (almost 90% in men). The probability of healthy patients being correctly identified is a bit lower, with values around 80%.

6. Conclusions

6.1 Study-related Conclusions

In first instance, the complexity of the database and the lineal dependences between its variables has proven to be very important. Most of the variables of every group (fat/lean variables, bone density measurements, etc) are strongly related one-another, so implementing techniques of dimensionality reduction is strongly recommended and useful.

Along with the lines of the first conclusion, the assumption of normality is delicate, as even when studied by genders variables show strong deviations from the gaussian distribution. This effect could be partially explained by the dual nature of our data (routine control analysis mixed with medical prescription ones).

Secondly, there is strong evidence to affirm that gender has an important role in this database (based not only on the results of the tests, but also on the biological background available in literature). Men and women show different mass values, fat distribution and bone density measurements, and therefore the data must be studied separately for every gender. The only variables that do not show gender differences are the T-scores, results that could be expected given that these are not observed values but standard deviations from healthy population.

Correlations inside every group of variables are evident if we observe at the correlation plots and graphical models. However, correlations between groups are less significant. This can not only be seen on the graphics, but also on the different machine learning models that try to predict a response variable (minTscore, Tscore_3cat or Tscore_2cat) without the variables used to calculate it (i.e., bone density measurements). If correlations between bone variables and fat/mass variables were stronger, we would expect better overall performances in the predictive models.

An important conclusion that emerges from the graphical models is that vertebrae variables seem to have little or no impact over the minTscore variable and, by extension, on osteoporosis/osteopenia diagnosis. As seen in the Gaussian Graphical Models (on both genders), only femoral variables have a direct relationship with the response variable. Further studies could deepen in this observation, as it could potentially reduce DEXA analysis costs and equipment dimensions (i.e., measuring the femoral bone vs measuring the whole skeleton). Also, scan repeatability might improve if vertebrae variables could be skipped: this region is problematic because body positioning inside the scanning machine is hard, specially if the patient suffers from kyphosis.

Tested machine learning algorithms performed significantly worse when bone variables were excluded from the analysis. None of the designed models has been capable of differentiating between osteoporosis and osteopenia: all models

trying to classify over 3 categories (healthy, osteopenia, osteoporosis) failed in their purpose. Not even balancing techniques (SNOTE) could improve the results. However, when joining the later categories in a single one (healthy vs diseased), model performances improved notoriously.

Random forests performed relatively poorly, even when classifying over only two categories. The other machine learning approaches reported better classificatory results, being the k-NN the best of the approaches. With the 15-NN model we achieved classificatory results with a sensitivity of the 80% (88% in men) and a specificity of the 80% (positive class: diseased). The results prove that muscular variables contain (somehow) enough information to determine the bone healthiness of a patient, but not enough to determine the degree of disease. Unfortunately, k-NN and SVM algorithms are a “black box” and the decisions/learnings made over the data to classify new observations will remain unknown.

We are safe to assure that relationships between fat/lean variables and the presence of osteoporosis/osteopenia exist. Further studies could deepen in this statement and determine if this relationships are causation or just correlation.

Another interesting work that could be conducted in the future is analyzing the evolution of every patient in time. Since DEXA scans are conducted periodically, it would be interesting to see how well the designed classification models perform at different temporal stages, and study if the correlations between variables change over time.

6.2 Personal growth

In the process of performing this study, I deepened in my knowledge about data management, study, and interpretation. I learned about the complexity that arises from working with a real dataset, and discovered techniques to manage missing data, outliers and typos. I learned the importance of implementing dimensionality reduction techniques (PCA) to manage correlated data, along with the potential of the new methods of both data observation (graphical models) and prediction (random forests, SVM, k-NN).

This work offered me the opportunity of experimenting with the theorical concepts studied in the master’s degree in the fields of biology, regression, modelling and machine learning. I also learned some traceability skills, useful to achieve result replication and workflow understanding.

This study also allowed me to improve my skills in the art of coding with R, along with implementing a dynamic code in Rmarkdown that allowed for fast replication of the analysis, adaptation to new observations and improvements in the code. With this study I found new R packages that will be of great utility in a future and,

more importantly, I discovered where and how I can search for both contrasted information and trustful statistical resources.

Conducting a study that aimed to shed light into a real scientific problem has been a great motivation, and even though the results of the modeling might not be good enough for medical application I feel happy with them. In them I see reflected the improvement of my statistical skills.

The continuous communication with Nuria Perez has been essential to achieve the milestones and objectives in time. She greatly helped me understand complex concepts and oriented my work when I walked into dead ends.

Some of the initial objectives could not be achieved due to a lack of time. I am mainly referring to the extension of the study to the lipodystrophy and low muscle mass comorbidities. Instead of repeating the analysis over the new response variables, deepening on the osteoporosis analysis has been preferred, leaving the other studies to a future work.

Overall, I performed accordingly to the calendar. Punctual deviations from the expected date limits have been solved reducing the complexity of the procedures. The most critical step has been to take the decision of not studying the other two comorbidities, which let me take more time to analyze osteoporosis in greater detail.

The rest of the milestones have been accomplished in time, as it has been reflected in the monitoring reports.

Conducting this final project helped me learn and practice rather intermediate/advanced statistical methods, but most importantly helped me realize that there is much more that what we can see. Databases hide a lot of information that is not visible at plain sight, and deepening into them to discover what they can offer is fascinating.

I wonder how many questions could be solved with the data that we already have, if we only knew how to look at it appropriately.

7. Glossary

DEXA: is an acronym for Dual-Energy X-ray Absorptiometry and refers to a common analytic technique to determine the levels of mineral density in the bones of a patient. Along with the bone measurements, DEXA scans also retrieve muscular and fat tissue values at various body regions.

T-test: T-test (or T-student test) is any statistical test that is guided by a statistical parameter that follows a Student's t-distribution under the null hypothesis. One of its more common uses (and how it has been used in this study) is to compare if the means of two populations are significantly different from each other (two sample T-test). In this study, the test compared variable means for every gender.

Shapiro-Wilk test: this is one of the most common tests to determine if a variable is normally distributed. In this test, the null hypothesis determines that the provided observations have been taken from a normally distributed population. Therefore, p-values below the significance level (alpha) suggest a low probability of the null hypothesis being true, leading us to reject the assumption of normality.

PCA: the best scenario to perform a study is when the variables on which we work are orthogonal, i.e., they show no correlation. This is a seldom scenario, especially in biological sciences. Principal Component Analysis is a technique to study a set of variables and find subsets of them that seek this least correlation, in what is called "dimensionality reduction". The variables given by the PCA (called Principal Components) are linear combinations of the original data, and ideally show less correlation than the original variables.

Machine learning: this concept is used to refer a group of algorithms that are capable of autonomously learn from our dataset and predict future behavior. These algorithms can detect complex patterns in our dataset that might be left unnoticed with classical regression methods, bringing opportunities to deeply study large and uniform databases. Graphical models, random forests, Support Vector Machines and k-Nearest Neighbors are (among others) machine learning-based analytical tools.

Graphical models: Gaussian and Mixed Graphical models (GGMs and MGMs, respectively) are probabilistic models that represent via dependency graphics the variables in our dataset (as nodes) and the relationships between them (as edges). In a simplistic way, the MGMs could be seen as an extension of the GGMs, designed to incorporate discrete and factorial variables.

Random forest: is a regression and classification tool, based on machine learning. This model combines the base principles of bagging (bootstrap aggregating) with decision trees. After a forest (ensemble of decision trees, each using a few of the original variables) is generated over multiple subsets of the

data, the model combines the outputs of every single tree to get a consensus prediction.

ROC / AUC: Receiver Operating Characteristic is a curve that represents the sensitivity over the specificity of a classifying model, giving us information about how good the model is at distinguishing between classes. The AUC is the area under said curve, with values ranging from 0.5 (curve following the diagonal) to 1 (curve converging with the upper-left vertex of the graphic). AUC values closer to 1 represent better classifying models. AUC values below 0.5 would represent models that are reciprocating classes.

OOB error: Out-of-Bag error estimates the prediction error in random forest and other bagging models. In said models, multiple subsets are taken over the original database (with replacement, called bagging or bootstrap aggregating). Therefore, every observation used to train the model is actually used only in “n” of the total decision trees calculated. Said observation will be predicted for all the models that did not use it in the training process, and the majority vote will classify it. The process is then repeated with all the observations, and the number of correct classifications gives us the OOB.

SVM: Support Vector Machines use multiple dimensions to create hyperplanes that correctly separate the groups in our dataset. Although they can be used for both regression and classification, only the later has been used in this study (which is, in fact, their most popular use). The most appealing trait about SVM is that they can not only model linear relationships, but also more complex situations (called “kernel functions”).

K-NN: k-Nearest Neighbors is a classification algorithm based on a common principle of machine learning: “similar observations are likely to have similar properties”. Based on this premise, the k-NN model evaluates new observations and assigns them to a category, based on the known categories of the “k” more likely observations in which the model has been trained.

Accuracy: percentage of observations correctly classified.

Kappa: correction of the accuracy value accounting for random accuracy. Especially relevant in scenarios where one event is much likely to happen than the other, and thus by simply classifying all individuals to the majority group we would achieve great values of accuracy.

Sensitivity: or true positive ratio, number of true positives detected divided by all positives (positives detected + positives classified as negative). A sensor of how well the model can detect a positive case.

Specificity: or true negative ratio, number of true negatives detected divided by all negatives (negatives detected + negatives classified as positive). A sensor of how reliable a positive value is.

8. Bibliography

Abdalla, P. P., Silva, A. M., Venturini, A. C. R., Santos, A. P. dos, Carvalho, A. dos S., Siqueira, V. A. A. A., … Machado, D. R. L. (2020). Cut-off points of appendicular lean soft tissue for identifying sarcopenia in the older adults in Brazil: a cross-sectional study. *Nutrición Hospitalaria*, *37*(2), 306–312.

Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H. J., & Zacharias, H. U. (2020). Gaussian and Mixed Graphical Models as (multi-)omics data analysis tools. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1863*(6), 194418. https://doi.org/10.1016/j.bbagrm.2019.194418

Beraldo, R. A., Vassimon, H. S., Aragon, D. C., Navarro, A. M., Albuquerque De Paula, F. J., & Foss-Freitas, M. C. (2015). Proposed ratios and cutoffs for the assessment of lipodystrophy in HIV-seropositive individuals. *European Journal of Clinical Nutrition*, *69*(2), 274–278. https://doi.org/10.1038/ejcn.2014.149

Bhushan, N., Mohnert, F., Sloot, D., Jans, L., Albers, C., & Steg, L. (2019). Using a Gaussian graphical model to explore relationships between items and variables in environmental psychology research. *Frontiers in Psychology*, *10*(MAY), 1–12. https://doi.org/10.3389/fpsyg.2019.01050

Compston, J. (2016). HIV infection and bone disease. *Journal of Internal Medicine*, *280*(4), 350–358. https://doi.org/10.1111/joim.12520

Cruz-Jentoft, A. J., Bahat, G., Bauer, J., Boirie, Y., Bruyère, O., Cederholm, T., … Schols, J. (2019). Sarcopenia: Revised European consensus on definition and diagnosis. *Age and Ageing*, *48*(1), 16–31. https://doi.org/10.1093/ageing/afy169

Doroudinia, A., & Colletti, P. M. (2015). Bone Mineral Measurements. *Clinical Nuclear Medicine*, *40*(8), 647–657. https://doi.org/10.1097/RLU.0000000000000860

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Finnerty, F., Walker-Bone, K., & Tariq, S. (2017). Osteoporosis in postmenopausal women living with HIV. *Maturitas*, *95*, 50–54. https://doi.org/10.1016/j.maturitas.2016.10.015

Freitas, P., Santos, A. C., Carvalho, D., Pereira, J., Marques, R., Martinez, E., … Medina, J. L. (2010). Fat Mass Ratio: An Objective Tool to Define Lipodystrophy in HIV-Infected Patients Under Antiretroviral Therapy. *Journal of Clinical Densitometry*, *13*(2), 197–203. https://doi.org/10.1016/j.jocd.2010.01.005

Guzman, N., & Vijayan, V. (2020). *HIV-associated Lipodystrophy*. StatPearls Publishing, Treasure Island (FL). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK493183/

Karastergiou, K., Smith, S. R., Greenberg, A. S., & Fried, S. K. (2012). Sex differences in human adipose tissues - The biology of pear shape. *Biology of Sex Differences*, *3*(13), 1–12. https://doi.org/10.1186/2042-6410-3-13

Lantz, B. (2015). *Machine Learning with R* (2nd ed.). Packt Publishing.

Liu, A. Y., Vittinghoff, E., Sellmeyer, D. E., Irvin, R., Mulligan, K., Mayer, K., … Buchbinder, S. P. (2011). Bone mineral density in HIV-negative men participating in a tenofovir pre-exposure prophylaxis randomized clinical trial in San Francisco. *PLoS ONE*, *6(8):e2368*. https://doi.org/10.1371/journal.pone.0023688

Malmstrom, T. K., Miller, D. K., Herning, M. M., & Morley, J. E. (2013). Low appendicular skeletal muscle mass (ASM) with limited mobility and poor health outcomes in middle-aged African Americans. *Journal of Cachexia, Sarcopenia and Muscle*, *4*(3), 179–186.

https://doi.org/10.1007/s13539-013-0106-x

Negredo, E., Domingo, P., Gutiérrez, F., Galindo, M. J., Knobel, H., Lozano, F., … Estrada, V. (2018). Executive summary of the consensus document on osteoporosis in HIV-infected individuals. *Enfermedades Infecciosas y Microbiologia Clinica*, *36*(5), 312–314. https://doi.org/10.1016/j.eimc.2017.03.010

Oliveira, V. H. F., Borsari, A. L., Webel, A. R., Erlandson, K. M., & Deminice, R. (2020). Sarcopenia in people living with the Human Immunodeficiency Virus: a systematic review and meta-analysis. *European Journal of Clinical Nutrition*, *74*(7), 1009–1021. https://doi.org/10.1038/s41430-020-0637-0

Oursler, K. K., Iranmanesh, A., Jain, C., Birkett, K. L., Briggs, B. C., Garner, D. C., … Ryan, A. S. (2020). Short Communication: Low Muscle Mass Is Associated with Osteoporosis in Older Adults Living with HIV. *AIDS Research and Human Retroviruses*, *36*(4), 300–302. https://doi.org/10.1089/aid.2019.0207

Premaor, M. O., & Compston, J. E. (2020). People living with HIV and fracture risk. *Osteoporosis International*, *31*(9), 1633–1644. https://doi.org/10.1007/s00198-020-05350-y

Royston, P. (1995). A Remark on Algorithm AS 181: The W-test for Normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *44*(4), 547–551.

Sedgewick, A. J., Buschur, K., Shi, I., Ramsey, J. D., Raghu, V. K., Manatakis, D. V., … Benos, P. V. (2019). Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics*, *35*(7), 1204–1212. https://doi.org/10.1093/bioinformatics/bty769

Shafiee, G., Ostovar, A., Heshmat, R., Keshtkar, A. A., Sharifi, F., Shadman, Z., … Larijani, B. (2018). Alloimmunization in thalassemia patients: New insight for healthcare. *International Journal of Preventive Medicine*, *9*(25). https://doi.org/10.4103/ijpvm.IJPVM

Studenski, S. A., Peters, K. W., Alley, D. E., Cawthon, P. M., McLean, R. R., Harris, T. B., … Vassileva, M. T. (2014). The FNIH sarcopenia project: Rationale, study description, conference recommendations, and final estimates. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *69*(5), 547–558. https://doi.org/10.1093/gerona/glu010

UNAIDS. (2020). UNAIDS data 2020. Geneva, Switzerland: UNAIDS. Retrieved from https://www.unaids.org/sites/default/files/media_asset/2020_aids-data-book_en.pdf

Viana, J. U., Dias, J. M. D., Pereira, L. S. M., Silva, S. L. A. da, Hoelzle, L. F., & Dias, R. C. (2018). Pontos de corte alternativos para massa muscular apendicular para verificação da sarcopenia em idosos brasileiros: dados da Rede Fibra - Belo Horizonte/Brasil. *Fisioterapia e Pesquisa*, *25*(2), 166–172. https://doi.org/10.1590/1809-2950/17533725022018

Williams, C. (2020). How to create a correlation matrix with too many variables in R. Retrieved October 10, 2020, from https://towardsdatascience.com/how-to-create-a-correlation-matri+x-with-too-many-variables-309cc0c0a57

9. Annexes

Annex 1: R packages and versions

Package	Version
car	3.0-10
carData	3.0-4
caret	6.0-86
corrplot	0.84
devtools	2.2.1
dplyr	0.8.3
factoextra	1.0.7
ggbiplot	0.55
ggplot2	3.3.2
gplots	3.0.1.1
gridExtra	2.3
haven	2.2.0
knitr	1.26
lattice	0.20-38
plyr	1.8.4
randomForest	4.6-14
randomForestExplainer	0.10.1
readxl	1.3.1
ROCR	1.0-7
rstudioapi	0.1
scales	1.0.0
usethis	1.5.1
xfun	0.11
pcalg	2.7-0
qgraph	1.6.5
ggm	2.5
Rgraphviz	2.28.0
RColorBrewer	1.1-2
mgm	1.2-10
UBL	0.0.6
kernlab	0.9-29
class	7.3-15
R software	**3.6.1 (2019-07-05)**
R studio	**1.2.5019**

Annex 2: Summary tables

Table 7: Summary of the deleted samples

Variable	Min	25%	Mean	75%	Max	SD
Age	29	40	46.24	50	67	8.35
Height	1.45	1.62	1.7	1.78	1.84	0.1
Weight	41	57.16	66.67	74.72	109.5	13.78
RAFp	4.3	7	16.13	22.5	36.5	8.44
RAFg	136	291	564.5	743	1257	319.9
RALg	1490	1943	2852	3462	4909	868.9
LAFp	4.4	7.1	16.59	22.6	35.9	9.02
LAFg	140	288	574.7	798	1392	345.1
LALg	1270	2004	2797	3394	5386	878.3
BothAFp	4.4	7.1	16.37	22.8	36.2	8.7
BothAFg	276	559	1140	1579	2580	662.9
BothALg	2760	3986	5649	6882	10296	1739
RLFp	4.2	10.5	17.79	23.3	39.6	9.53
RLFg	380	1106	1832	2623	6006	1161
RLLg	4309	6437	8035	9280	14476	2035
LLFp	4.2	10.7	17.79	23.4	40.2	9.56
LLFg	369	1113	1863	2801	6099	1182
LLLg	4288	6411	8030	9198	14428	2049
BothLFp	4.2	10.4	17.78	23.4	39.9	9.55
BothLFg	749	2258	3715	5559	12105	2330
BothLLg	8597	13002	16067	18299	28904	4075
TFp	7.2	15.8	24.83	33.1	42.6	9.84
TFg	2364	4873	8724	12236	18050	4525
TLg	16628	20856	24599	28068	40608	4884
TotalFp	5.5	14.1	20.97	28.2	35.8	8.39
TotalFg	3536	8286	13966	20096	27462	6736
TotalLg	28630	40290	50093	57324	84066	11008
L1BMD	0.73	0.95	1.04	1.15	1.44	0.16
L1T	-3.6	-1.75	-0.92	-0.1	2.3	1.27
L1Z	-3	-1.33	-0.56	-0.08	1.7	1.11
L2BMD	0.81	1.02	1.14	1.24	1.67	0.18
L2T	-3.4	-1.75	-0.77	0	3.6	1.51
L2Z	-2.8	-1.4	-0.39	0.3	3	1.33
L3BMD	0.78	1.03	1.15	1.24	1.76	0.19
L3T	-3.8	-1.65	-0.61	0.15	4.4	1.59
L3Z	-3.2	-1.35	-0.26	0.65	3.8	1.49
L4BMD	0.85	1.01	1.13	1.27	1.68	0.19
L4T	-3.3	-1.9	-0.76	0.3	3.7	1.6

L4Z	-3	-1.7	-0.43	0.68	3.5	1.58
L1L4BMD	0.8	1	1.12	1.22	1.64	0.18
L1L4T	-3.5	-1.72	-0.76	0.1	3.5	1.45
L1L4Z	-2.9	-1.5	-0.39	0.25	2.9	1.33
L2L4BMD	0.82	1.01	1.14	1.26	1.7	0.18
L2L4T	-3.5	-1.75	-0.71	0.1	3.9	1.5
L2L4Z	-2.9	-1.4	-0.32	0.6	3.3	1.45
NeckFBMD	0.73	0.86	0.94	1.02	1.27	0.12
NeckFT	-2.6	-1.4	-0.82	-0.3	1.5	0.93
NeckFZ	-2.1	-0.8	-0.21	0.4	2	0.88
WardsBMD	0.56	0.7	0.78	0.86	1.04	0.14
WardsT	-3.1	-1.9	-1.23	-0.55	0.6	1.04
WardsZ	-2.4	-1.1	-0.34	0.52	1.4	0.95
TrochBMD	0.51	0.71	0.77	0.86	1.05	0.13
TrochT	-3.8	-1.83	-1.07	-0.27	1.1	1.05
TrochZ	-2.6	-1.4	-0.71	-0.05	1.1	0.92
TotalFBMD	0.7	0.88	0.97	1.06	1.3	0.13
TotalFT	-2.7	-1.4	-0.77	-0.1	1.7	0.95
TotalFZ	-2.1	-0.9	-0.28	0.2	2.1	0.85
BMI	17.51	20.91	23.01	25.09	33.06	3.43
FMI	1.14	2.89	4.88	7.19	8.73	2.28
FFMI	12.23	14.73	17.16	18.52	25.38	2.63
Apendicularleanmas	4.85	6.06	7.42	8.18	11.83	1.41
FMR	0.48	1.13	1.7	2.13	4.34	0.83
FTrunkgFLegsg	0.76	1.75	2.92	3.75	9.21	1.66
Indexdistributionfat	0.59	1.33	2.09	2.61	5.42	1.02
FtrunkpFlimbsp	0.36	0.55	0.67	0.74	1.18	0.17
FtrunkgFtotalg	0.37	0.55	0.63	0.7	0.82	0.11
FLegsgFtotalg	0.09	0.18	0.26	0.32	0.48	0.1
FlimbsgFtotalg	0.15	0.27	0.34	0.41	0.63	0.1
LLegFgBMI	16.81	46.8	81.15	108.2	232.1	47.94
LLegFpBMI	0.2	0.48	0.79	1.04	1.85	0.44
minTscore	-3.8	-2.1	-1.45	-0.8	1.5	1.04
TotalBMD	0.81	1.09	1.16	1.22	1.75	0.15
Disease_age	1.26	8.13	14.7	21.1	31.2	8.44

Table 8: Summary of the variables

Variable	Min	25%	Mean	75%	Max	SD
Age	17	39	45.92	53	81	10.57
Height	1.4	1.64	1.7	1.76	1.93	0.09
Weight	34.6	60.56	69.63	78.04	120.5	12.58
RAFp	3.7	10	18.49	24.87	65.6	10.67
RAFg	49	356.2	701.4	949.7	4374	459.9
RALg	1125	2351	2889	3408	9316	799.8
LAFp	2.2	9.93	18.59	24.82	63	10.74
LAFg	45	352.2	693.3	926.7	3532	459.8
LALg	299	2298	2819	3321	6659	764.6
BothAFp	3.6	10	18.56	24.9	64.4	10.7
BothAFg	95	697.2	1392	1877	7907	919.7
BothALg	2271	4656	5707	6710	13317	1534
RLFp	3.8	11.65	20.54	27.6	63	10.98
RLFg	244	1179	2271	3035	12570	1444
RLLg	3719	6848	8050	9266	13889	1746
LLFp	3.8	11.6	20.48	27.48	63.8	10.96
LLFg	239	1179	2269	3051	12570	1436
LLLg	3815	6738	8053	9267	13873	1755
BothLFp	3.8	11.6	20.47	27.48	63.3	10.93
BothLFg	484	2358	4534	6052	25140	2870
BothLLg	7776	13600	16102	18521	27395	3477
TFp	4.5	21.02	28.71	36.48	59.4	10.6
TFg	1006	6595	10557	13857	34163	5096
TLg	12471	21259	24332	27355	59172	4644
TotalFp	4.2	16.72	24.14	30.3	56.6	9.62
TotalFg	184	10736	16794	21401	52915	8118
TotalLg	25056	43454	49891	56342	88914	9687
L1BMD	0.58	0.94	1.05	1.14	1.79	0.15
L1T	-4.5	-1.7	-0.85	-0.1	3.4	1.27
L1Z	-3.7	-1.4	-0.56	0.2	4.6	1.21
L2BMD	0.12	1.01	1.12	1.23	1.68	0.16
L2T	-5.5	-1.8	-0.92	0	3.5	1.34
L2Z	-4.8	-1.5	-0.62	0.2	4.1	1.29
L3BMD	0.34	1.02	1.13	1.23	1.71	0.17
L3T	-4.9	-1.8	-0.82	0.1	3.9	1.39
L3Z	-4.2	-1.5	-0.53	0.3	4.6	1.35
L4BMD	0.62	0.99	1.1	1.21	1.76	0.17
L4T	-4.8	-2	-1.06	-0.2	4.3	1.38
L4Z	-4.3	-1.7	-0.77	0.08	4.5	1.35
L1L4BMD	0.59	0.99	1.1	1.2	1.67	0.15

L1L4T	-4.9	-1.8	-0.9	-0.1	3.7	1.28
L1L4Z	-4	-1.4	-0.61	0.1	3.6	1.23
L2L4BMD	0.59	1.01	1.12	1.22	1.71	0.16
L2L4T	-5.1	-1.9	-0.94	-0.1	3.9	1.31
L2L4Z	-4.4	-1.5	-0.65	0.1	4	1.27
NeckFBMD	0.51	0.84	0.94	1.03	1.74	0.14
NeckFT	-4	-1.6	-0.85	-0.2	5.1	1.09
NeckFZ	-3.1	-0.9	-0.27	0.3	5.4	0.94
WardsBMD	0.34	0.66	0.78	0.88	1.79	0.16
WardsT	-4.5	-2.2	-1.32	-0.5	6.4	1.27
WardsZ	-3.4	-1.2	-0.5	0.1	6.9	1.08
TrochBMD	0.35	0.7	0.79	0.87	1.62	0.13
TrochT	-4.3	-1.8	-0.97	-0.2	6.2	1.17
TrochZ	-3.8	-1.4	-0.65	0	6.2	1.07
TotalFBMD	0.3	0.88	0.97	1.06	1.67	0.14
TotalFT	-4	-1.5	-0.73	0	4.1	1.1
TotalFZ	-3.3	-1	-0.3	0.3	5.2	0.99
BMI	14.04	21.45	23.96	26.09	40.81	3.6
FMI	0.06	3.63	5.86	7.48	19.71	2.99
FFMI	10.71	15.49	17.08	18.54	27.85	2.33
Apendicularleanmas	4.12	6.59	7.45	8.26	11.77	1.23
FMR	0.47	1.12	1.66	2.02	8.88	0.79
FTrunkgFLegsg	0.5	1.72	2.87	3.6	16.4	1.66
Indexdistributionfat	0.36	1.37	2.08	2.56	8.11	1
FtrunkpFlimbsp	0.36	0.56	0.67	0.74	1.83	0.17
FtrunkgFtotalg	0.26	0.56	0.63	0.7	0.99	0.1
FLegsgFtotalg	0.05	0.19	0.27	0.33	0.55	0.09
FlimbsgFtotalg	0.11	0.27	0.35	0.41	0.74	0.1
LLegFgBMI	14.59	50.59	92.14	125.4	344.3	50.34
LLegFpBMI	0.15	0.5	0.85	1.12	2.16	0.44
minTscore	-4.9	-2.2	-1.47	-0.8	2.4	1.04
TotalBMD	0.72	1.09	1.16	1.23	1.9	0.11
Disease_age	0.1	7.63	15.28	22.49	38.61	9

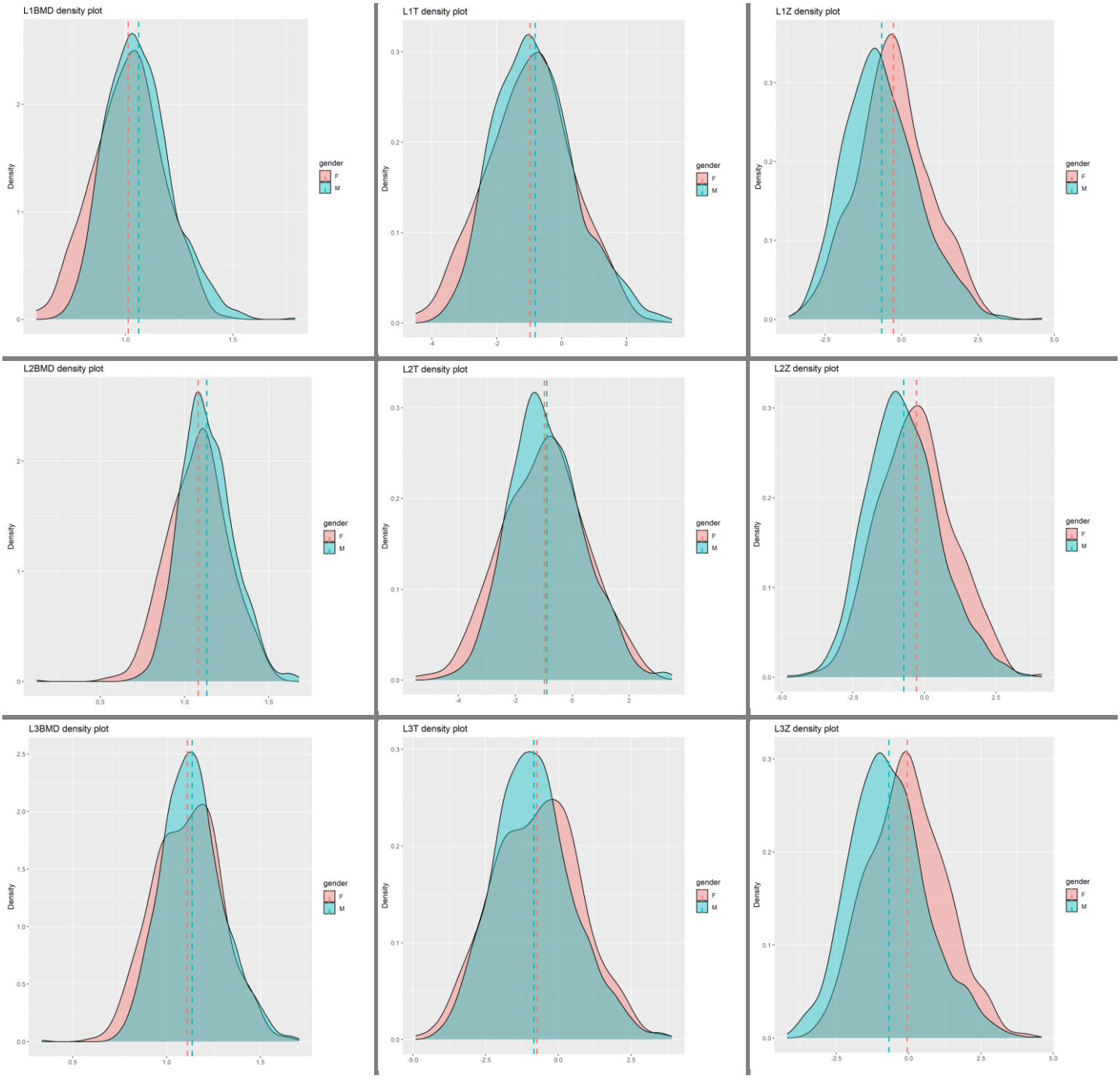
Table 9: Summary of the variables, by gender

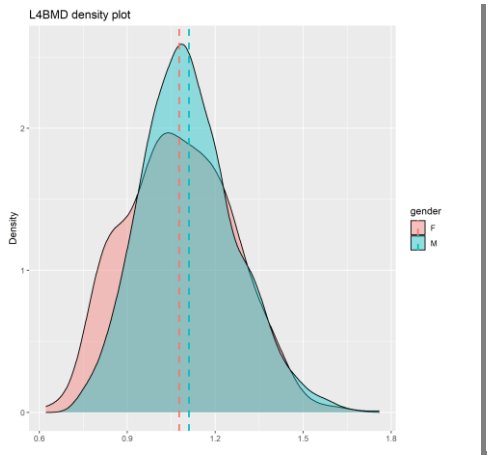
Variable	25% F	Mean F	75% F	SD F	25% M	Mean M	75% M	SD M
Age	41	46.91	52	9.55	38	45.6	53	10.85
Height	1.56	1.61	1.65	0.07	1.68	1.73	1.78	0.07
Weight	51.88	59.64	64.72	11.98	65.23	72.8	79.77	11.01
RAFp	20.98	28.84	37.52	11.55	8.4	15.2	20.5	7.93
RAFg	519.3	907.2	1143	581.1	323	636	877.2	392.4
RALg	1649	1890	2079	353.7	2794	3207	3539	620.1
LAFp	20.9	29.03	37.7	11.58	8.4	15.28	20.7	7.98
LAFg	499	900	1134	587.4	318	627.7	853.7	389.1
LALg	1606	1852	2042	358	2727	3127	3456	581
BothAFp	20.98	28.95	37.6	11.56	8.4	15.25	20.6	7.95
BothAFg	1010	1802	2279	1169	636.5	1262	1726	781.7
BothALg	3269	3742	4102	701.6	5518	6332	6988	1151
RLFp	23.03	31.22	38.9	11.41	10.2	17.14	22.7	8.36
RLFg	1869	3091	3802	1789	1063	2010	2735	1204
RLLg	5235	5911	6401	1083	7848	8730	9544	1314
LLFp	22.67	31.16	38.9	11.46	10.2	17.08	22.6	8.29
LLFg	1845	3075	3802	1748	1048	2013	2735	1215
LLLg	5259	5901	6394	1078	7818	8737	9560	1322
BothLFp	22.6	31.07	38.9	11.51	10.2	17.1	22.6	8.27
BothLFg	3700	6133	7608	3522	2110	4025	5455	2419
BothLLg	10480	11805	12839	2143	15726	17468	19107	2595
TFp	27.6	34.85	42.92	10.81	19.52	26.76	34.2	9.76
TFg	7002	11244	14292	5384	6462	10339	13694	4984
TLg	16872	18875	20486	2923	23582	26067	28060	3637
TotalFp	25.78	32.04	39.5	9.78	15.5	21.62	27.7	8.1
TotalFg	13354	19575	24071	9191	10000	15910	20580	7538
TotalLg	33768	37561	40262	5647	49071	53811	57825	7058
L1BMD	0.91	1.01	1.12	0.16	0.95	1.06	1.15	0.15
L1T	-1.9	-0.98	-0.1	1.32	-1.7	-0.81	-0.1	1.25
L1Z	-1.1	-0.27	0.43	1.18	-1.5	-0.65	0.1	1.2
L2BMD	0.95	1.08	1.19	0.17	1.03	1.13	1.23	0.16
L2T	-2.1	-0.98	0	1.46	-1.8	-0.89	-0.1	1.3
L2Z	-1.2	-0.28	0.6	1.3	-1.6	-0.73	0	1.26
L3BMD	0.98	1.11	1.24	0.18	1.02	1.14	1.23	0.16
L3T	-1.8	-0.74	0.3	1.48	-1.8	-0.84	-0.1	1.35
L3Z	-1	-0.05	0.9	1.31	-1.6	-0.69	0.1	1.33
L4BMD	0.95	1.08	1.21	0.18	1	1.11	1.21	0.16
L4T	-2.1	-1.03	0.1	1.49	-2	-1.07	-0.2	1.35
L4Z	-1.3	-0.34	0.5	1.34	-1.8	-0.91	-0.1	1.32
L1L4BMD	0.95	1.07	1.18	0.16	1.01	1.11	1.21	0.15

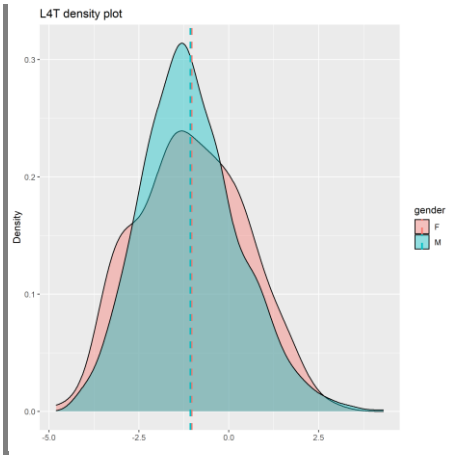
L1L4T	-1.9	-0.9	0.02	1.38	-1.8	-0.9	-0.1	1.24
L1L4Z	-1.1	-0.2	0.5	1.21	-1.5	-0.73	0	1.21
L2L4BMD	0.96	1.09	1.21	0.17	1.02	1.13	1.22	0.15
L2L4T	-2	-0.93	0.1	1.42	-1.8	-0.94	-0.1	1.28
L2L4Z	-1.12	-0.23	0.52	1.25	-1.6	-0.78	-0.1	1.25
NeckFBMD	0.79	0.89	0.98	0.14	0.86	0.95	1.04	0.14
NeckFT	-1.6	-0.79	0	1.2	-1.6	-0.87	-0.2	1.05
NeckFZ	-0.9	-0.24	0.3	0.99	-0.9	-0.29	0.3	0.92
WardsBMD	0.62	0.75	0.86	0.17	0.67	0.78	0.89	0.16
WardsT	-2.2	-1.25	-0.4	1.29	-2.2	-1.34	-0.6	1.26
WardsZ	-1.1	-0.36	0.3	1.08	-1.3	-0.55	0	1.07
TrochBMD	0.62	0.71	0.8	0.13	0.73	0.81	0.89	0.12
TrochT	-1.6	-0.74	0.02	1.21	-1.8	-1.04	-0.4	1.15
TrochZ	-1.1	-0.42	0.3	1.06	-1.4	-0.72	-0.1	1.06
TotalFBMD	0.8	0.91	1.01	0.15	0.9	0.99	1.08	0.14
TotalFT	-1.7	-0.73	0.1	1.25	-1.4	-0.73	-0.1	1.05
TotalFZ	-1.12	-0.38	0.3	1.07	-0.9	-0.27	0.3	0.96
BMI	20.41	23.16	25.4	4.52	21.86	24.21	26.18	3.22
FMI	5.23	7.65	9.76	3.64	3.39	5.3	6.9	2.5
FFMI	13.47	14.55	15.4	1.72	16.67	17.88	19.01	1.87
Apendicul.	5.48	6.01	6.37	0.83	7.27	7.9	8.52	0.96
FMR	0.9	1.26	1.39	0.69	1.24	1.79	2.2	0.78
FTrunkgFLegsg	1.29	2.23	2.64	1.58	1.9	3.08	3.95	1.64
Indexdistr.	1.07	1.59	1.84	0.79	1.54	2.23	2.77	1.01
FtrunkpFlimbsp	0.5	0.55	0.58	0.1	0.6	0.71	0.77	0.16
FtrunkgFtotalg	0.51	0.57	0.63	0.09	0.59	0.65	0.72	0.09
FLegsgFtotalg	0.24	0.32	0.39	0.1	0.18	0.25	0.31	0.08
FlimbsgFtotalg	0.34	0.4	0.47	0.09	0.26	0.33	0.39	0.09
LLegFgBMI	87.7	128	160.3	54.86	46.09	80.74	108.1	42.97
LLegFpBMI	1.03	1.34	1.64	0.42	0.44	0.7	0.91	0.31
minTscore	-2.2	-1.39	-0.5	1.15	-2.2	-1.5	-0.9	1.01
TotalBMD	1.03	1.1	1.17	0.11	1.11	1.18	1.24	0.1
Disease_age	13.21	18.54	24.54	8.02	6.56	14.25	21.22	9.05

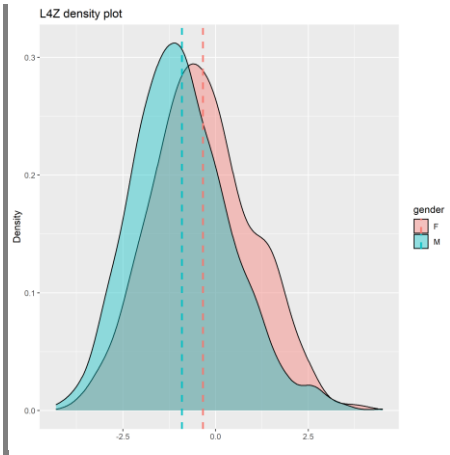
Annex 3: Density plots

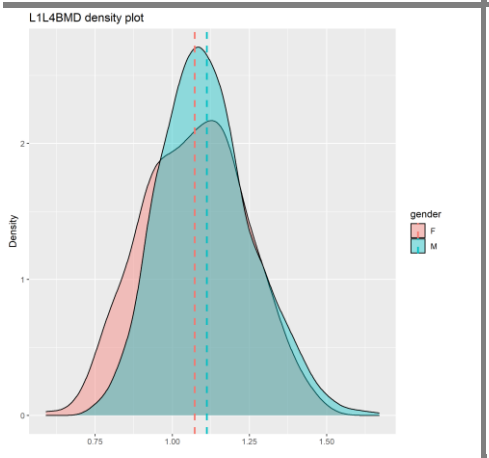
Figure 17: Density plots (bone variables). Left to right columns: BMI scores, T-scores and Z-scores.

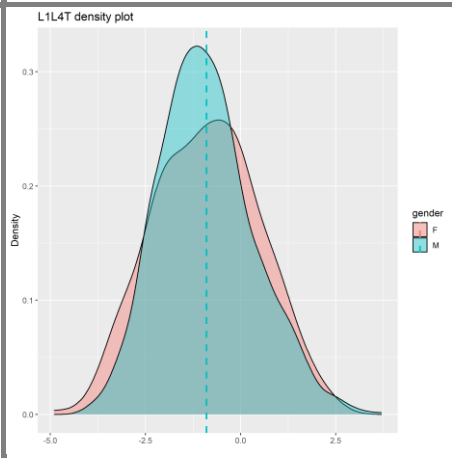


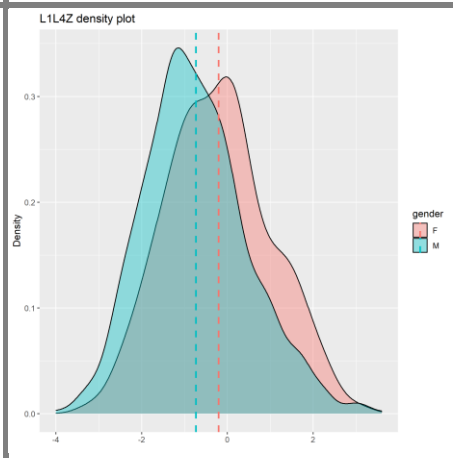

L4BMD density plot
Density
gender
F
M

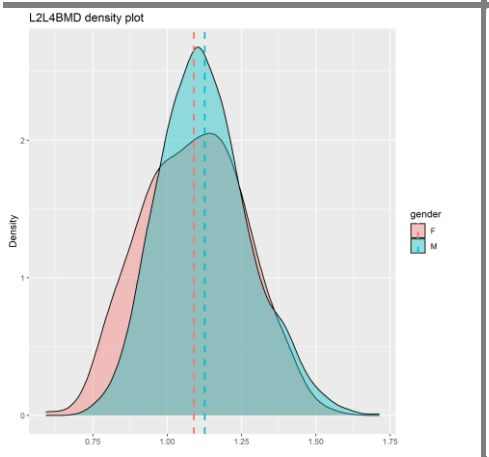

L4T density plot
Density
gender
F
M

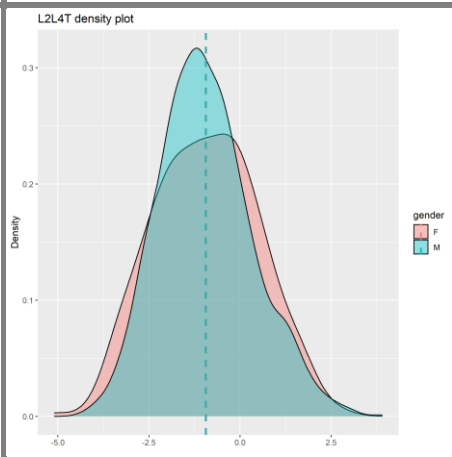

L4Z density plot
Density
gender
F
M

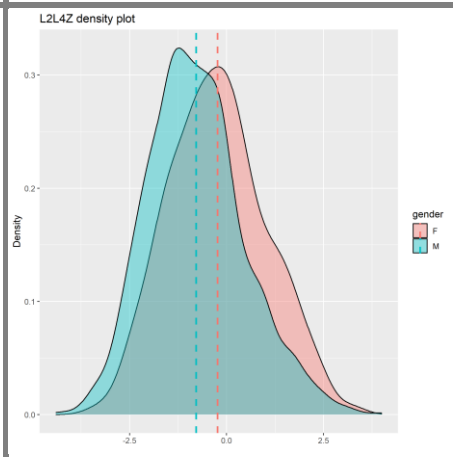

L1L4BMD density plot
Density
gender
F
M

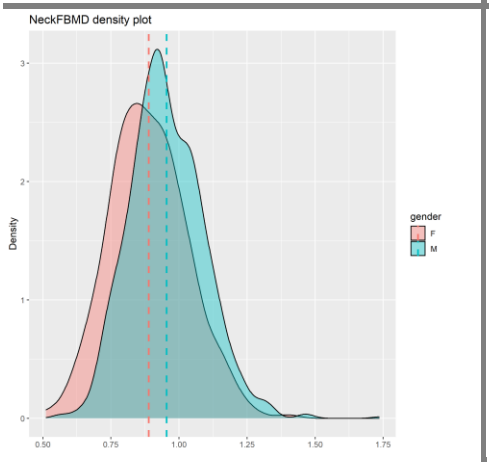

L1L4T density plot
Density
gender
F
M

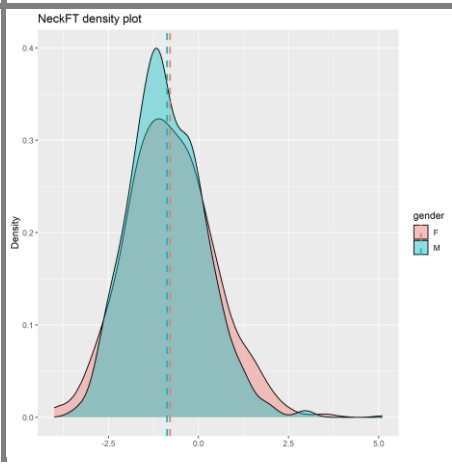

L1L4Z density plot
Density
gender
F
M

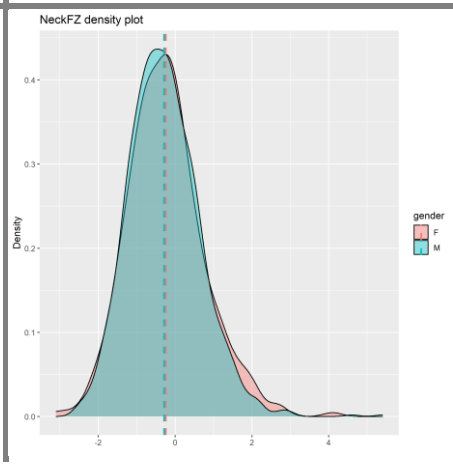

L2L4BMD density plot
Density
gender
F
M


L2L4T density plot
Density
gender
F
M


L2L4Z density plot
Density
gender
F
M


NeckFBMD density plot
Density
gender
F
M


NeckFT density plot
Density
gender
F
M


NeckFZ density plot
Density
gender
F
M

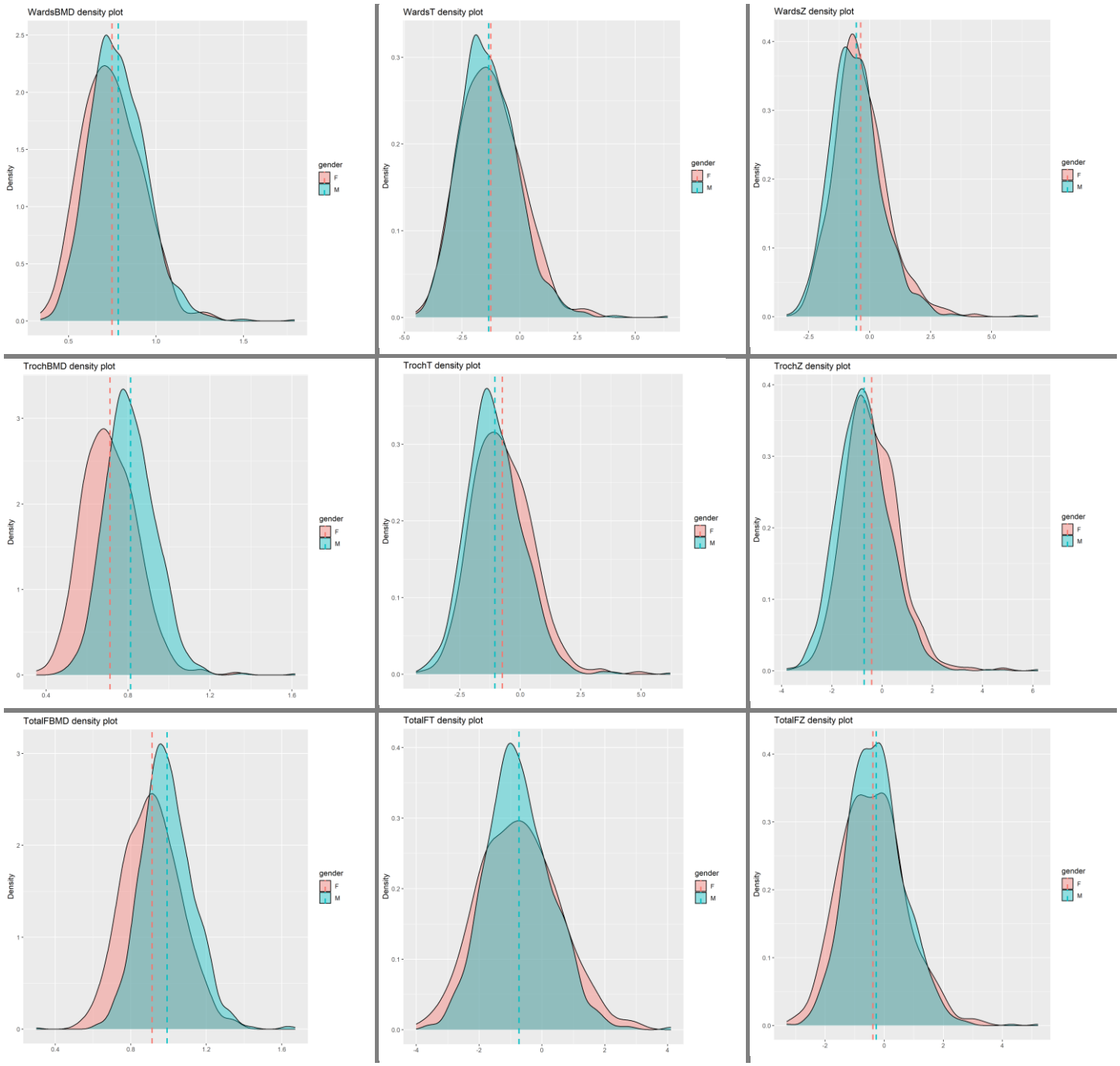

WardsBMD density plot
WardsT density plot
WardsZ density plot
TrochBMD density plot
TrochT density plot
TrochZ density plot
TotalFBMD density plot
TotalFT density plot
TotalFZ density plot
Density
gender
F
M

Figure 18: Density plots (fat variables). Columns represent the body side (left, right, both or total). The first two rows refer to upper limbs (in grams and percentages, respectively), while the last two rows refer to the legs.

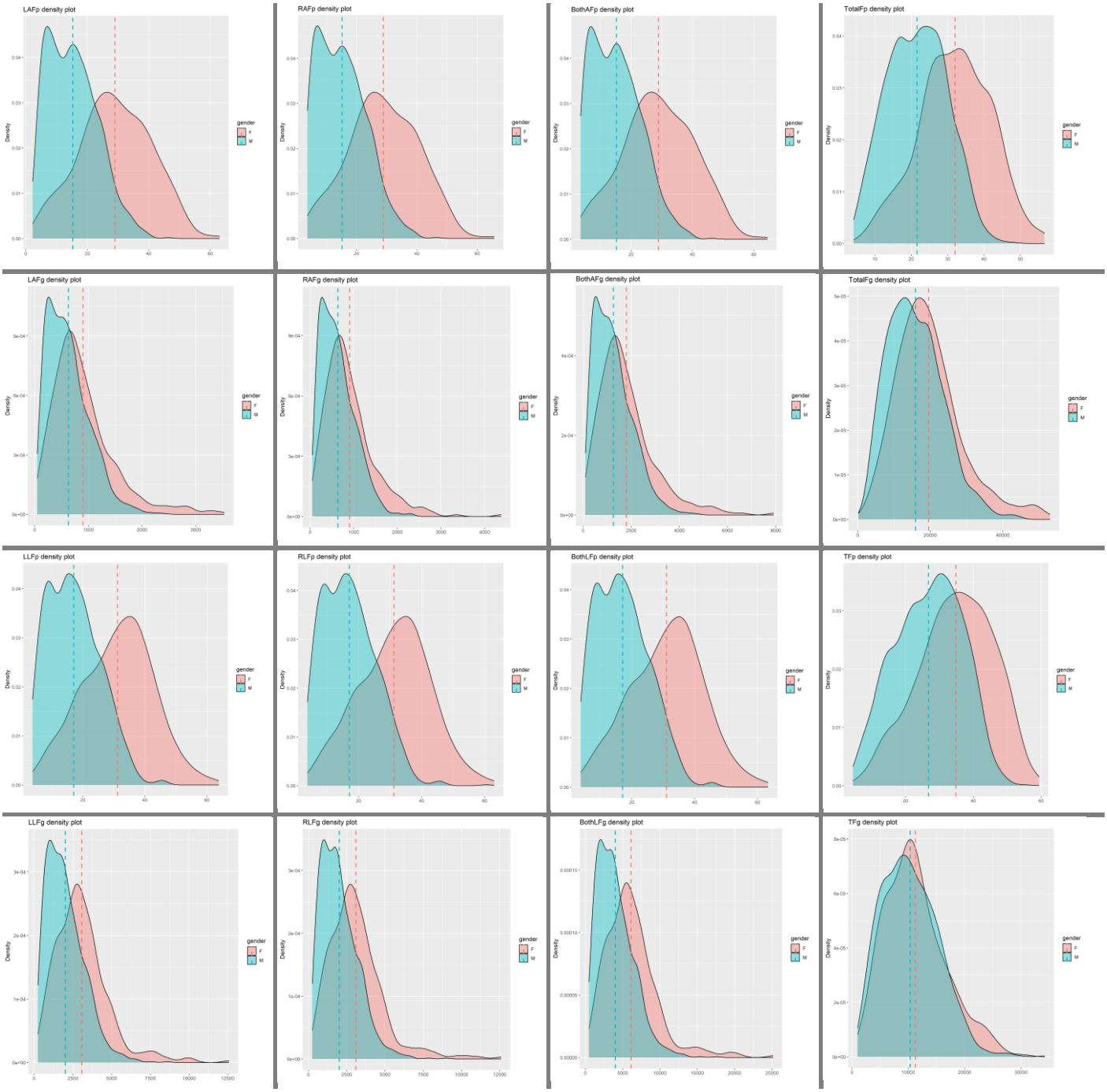


Figure 19: Density plots (lean variables). Columns represent the body side (left, right, both or total). The first row refers to upper limbs, while the second one refers to the legs.

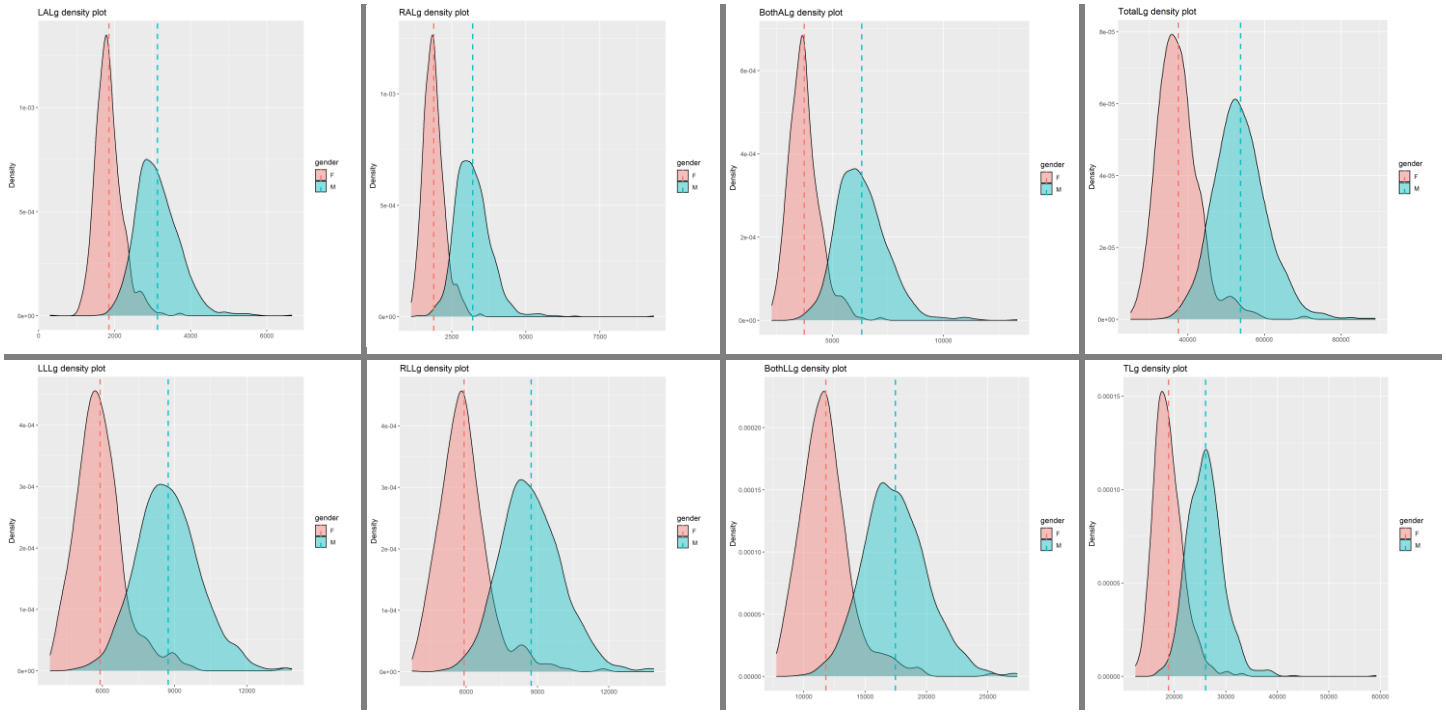
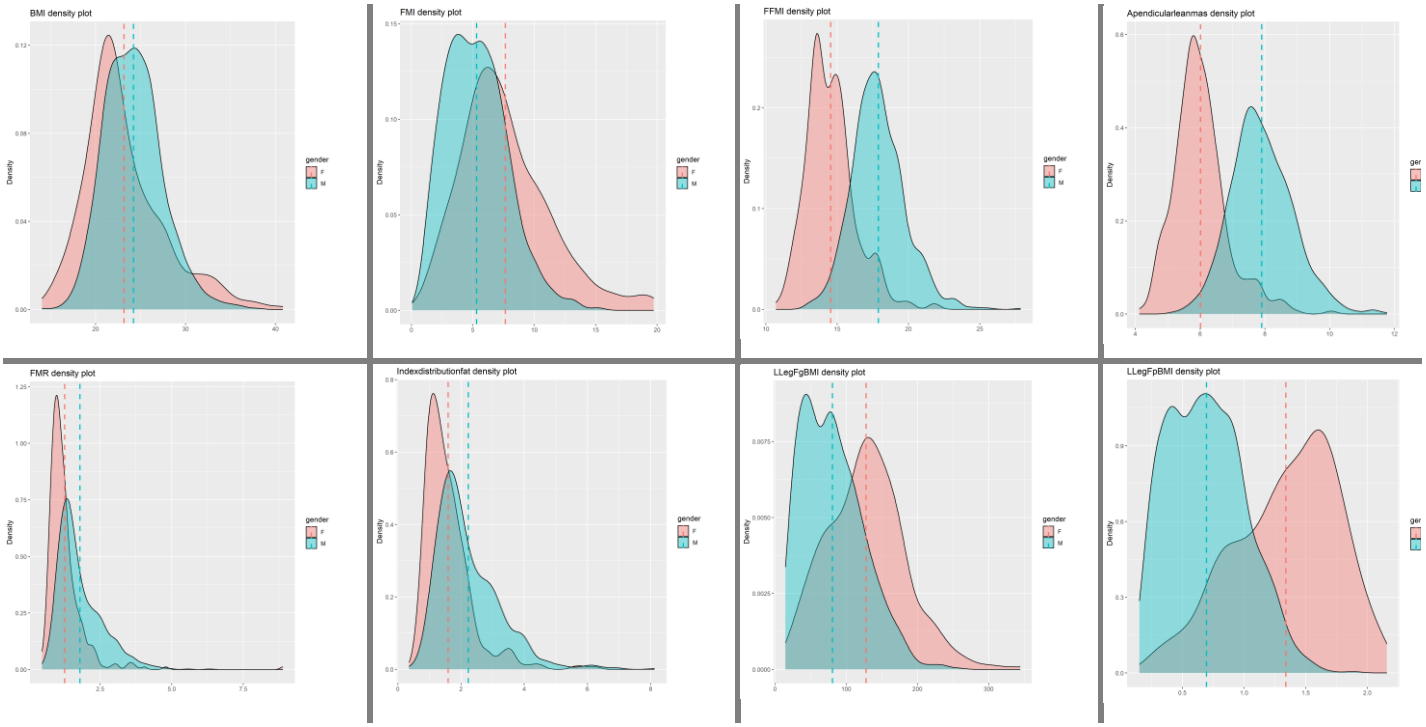
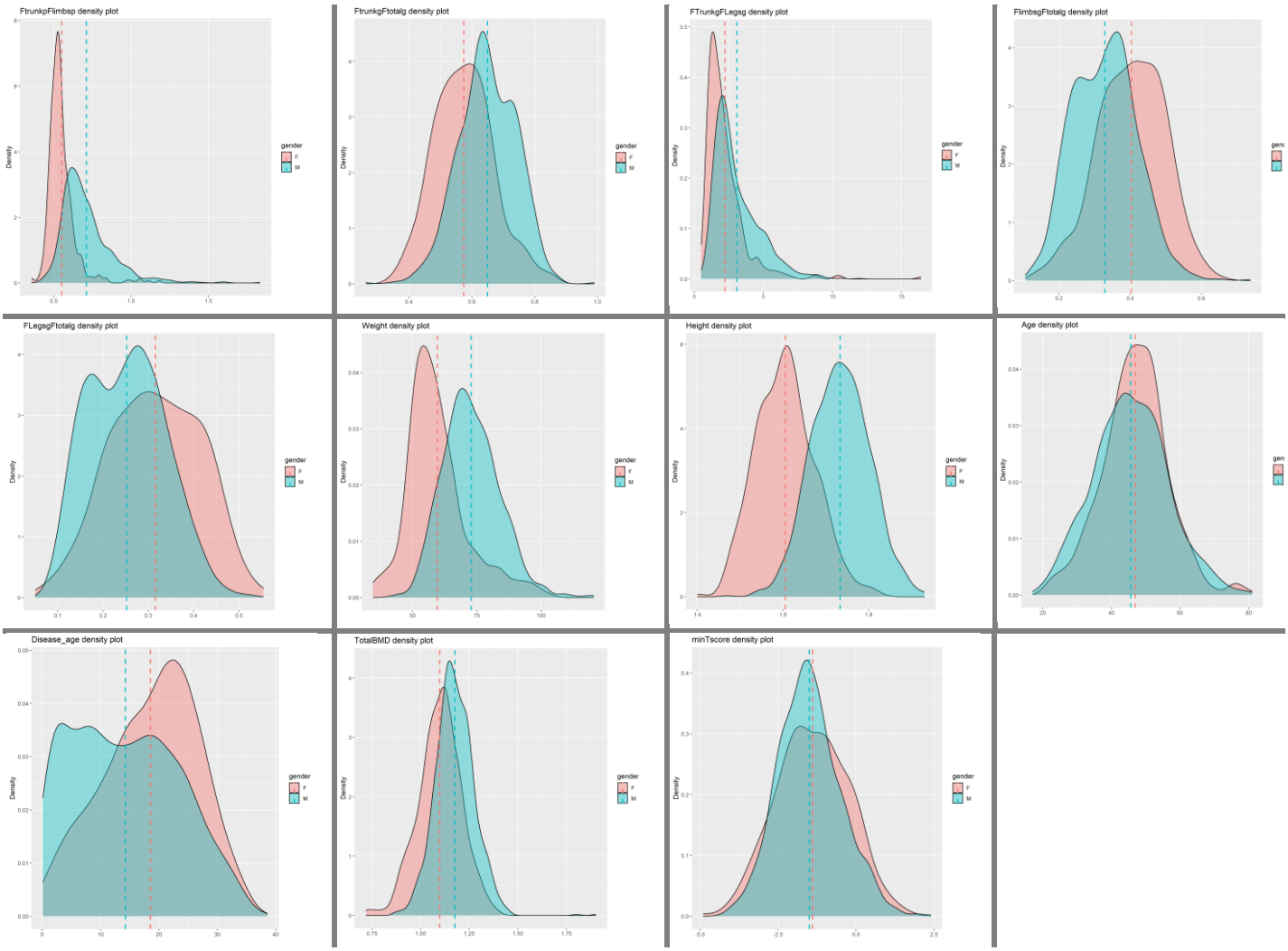


Figure 20: Density plots (summary variables).




FtrunkpFlimbsp density plot
FtrunkgFtotalg density plot
FTrunkgFLegsg density plot
FlimbsgFtotalg density plot
FLegsgFtotalg density plot
Weight density plot
Height density plot
Age density plot
Disease_age density plot
TotalBMD density plot
minTscore density plot
Density
gender
F
M

Annex 4: Directed Gaussian Graphical Models

Variable legend for the following graphics:

[1] Age
[2] Height
[3] Weight
[4] RAFp
[5] RAFg
[6] RALg
[7] LAFp
[8] LAFg
[9] LALg
[10] BothAFp
[11] BothAFg
[12] BothALg
[13] RLFp
[14] RLFg
[15] RLLg
[16] LLFp
[17] LLFg
[18] LLLg
[19] BothLFp
[20] BothLFg
[21] BothLLg
[22] TFp
[23] TFg
[24] TLg
[25] TotalFp
[26] TotalFg
[27] TotalLg
[28] L1BMD
[29] L1T
[30] L1Z
[31] L2BMD
[32] L2T
[33] L2Z
[34] L3BMD
[35] L3T
[36] L3Z
[37] L4BMD
[38] L4T
[39] L4Z
[40] L1L4BMD
[41] L1L4T
[42] L1L4Z
[43] L2L4BMD
[44] L2L4T
[45] L2L4Z
[46] NeckFBMD
[47] NeckFT
[48] NeckFZ
[49] WardsBMD
[50] WardsT
[51] WardsZ
[52] TrochBMD
[53] TrochT
[54] TrochZ
[55] TotalFBMD
[56] TotalFT
[57] TotalFZ
[58] BMI
[59] FMI
[60] FFMI
[61] Apendicularleanmas
[62] FMR
[63] FTrunkgFLegsg
[64] Indexdistributionfat
[65] FtrunkpFlimbsp
[66] FtrunkgFtotalg
[67] FLegsgFtotalg
[68] FlimbsgFtotalg
[69] LLegFgBMI
[70] LLegFpBMI
[71] minTscore
[72] TotalBMD
[73] Disease_age

Figure 21: Female (upper) and male (lower) directed GGM.