Universitat Oberta de Catalunya

uoc.edu

Machine learning methods for cross-sectional and longitudinal study of abnormal body fat distribution in HIV-infected individuals

Paola Fuentes Claramonte
Máster Universitario en Bioinformática y Bioestadística
Área 2 – Análisis de datos

Consultor/a: Nuria Pérez Álvarez
Profesor/a responsable de la asignatura: Marc Maceira Duch

Fecha Entrega: 5 enero 2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Machine learning methods for cross-sectional and longitudinal study of abnormal body fat distribution in HIV-infected individuals*
Nombre del autor:	*Paola Fuentes Claramonte*
Nombre del consultor/a:	*Nuria Pérez Álvarez*
Nombre del PRA:	*Marc Maceira Duch*
Fecha de entrega:	*01/2021*
Titulación:	*Máster Universitario en Bioestadística y Bioinformática*
Área del Trabajo Final:	*Área 2 – Análisis de datos*
Idioma del trabajo:	*Inglés*
Palabras clave	*DEXA/DXA, HIV, Machine Learning*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

La lipodistrofia es una alteración en la distribución de la grasa corporal asociada al VIH y su tratamiento farmacológico, que puede ser factor de riesgo para otros problemas de salud, por lo que su identificación y predicción pueden contribuir a mejorar la calidad de vida de estos pacientes.

El objetivo de este trabajo era aplicar métodos de *machine learning* (ML) a una base de datos real con medidas de densidad ósea, masa magra y masa grasa obtenidas mediante DEXA de una muestra de pacientes con infección por VIH, con medidas repetidas, buscando desarrollar herramientas para identificar la lipodistrofia y predecir su evolución.

Primero se estudió la estructura de los datos mediante métodos correlacionales y PCA, encontrando altas correlaciones entre variables, con 6 componentes principales que podían explicar más del 90% de la varianza original contenida en 58 variables.

Los modelos de ML mostraron, a nivel transversal, una clasificación muy precisa de la lipodistrofia si se incluían en el modelo variables cuantificando la masa grasa, pero un rendimiento pobre si se intentaba predecir la lipodistrofia a partir de otros tejidos. Para incorporar la estructura longitudinal, se utilizaron modelos lineales mixtos y una aproximación combinada (MEml, *Mixed Effects machine learning*). Ambas técnicas mostraron una buena capacidad predictiva. El método de MEml permite además predecir la lipodistrofia a nivel longitudinal.

Los resultados indican el potencial de los métodos de ML para la clasificación y predicción de las alteraciones en la distribución de los tejidos corporales en el

contexto de la infección por VIH.

Abstract (in English, 250 words or less):

Lipodystrophy is an alteration of body fat distribution associated to HIV and its pharmacological treatment, which can be a risk factor for other health problems, so its identification and prediction may contribute to improve the quality of life of these patients.

The goal of this work was to apply machine learning (ML) methods to a real dataset containing DXA-derived measures of bone mineral density, lean mass and fat mass from a sample of HIV-infected patients, with repeated measures, aiming to develop tools for identifying and predicting the evolution of lipodystrophy.

First, correlational methods and PCA were used to examine data structure, and results showed high correlations among variables, with 6 principal components explaining more than 90% of the original variance contained in 58 variables.

ML models showed, cross-sectionally, a very precise classification performance of lipodystrophy cases when variables quantifying fat mass or percentage were included in the models, but poor performance if prediction was based on other body tissues. To incorporate the longitudinal structure, linear mixed models and a combined approach (MEml, Mixed Effects machine learning) were used. Both methods showed good predictive capacity. MEml models allow, in addition, the longitudinal prediction of lipodystrophy.

Results highlight the potential of ML methods for classification and prediction of body tissue distribution alterations in the context of HIV infection.

Index

1. Introduction 7
1.1 Context and justification of the work 7
1.2 Project goals 9
1.3 Methodological approach 10
1.4 Work plan 11
1.5 Brief summary of the resulting products 13
1.6 Brief summary of the remaining chapters 14
2. Background 16
2.1 Lipodystrophy in the context of HIV infection 16
2.2 Methods to analyze highly correlated data 18
2.3 Machine learning methods for classification and prediction 20
3. Materials and Methods 24
3.1 Dataset 24
3.3 Exploration of the correlation structure 25
3.4 Data reduction: Principal Component Analysis 25
4. Results 31
4.1 Dataset description 31
4.2 Identification of body tissue distribution alterations 31
4.3 Correlation structure 32
4.4 Principal component analysis 34
4.5 Cross-sectional classification of fat mass distribution alterations 39
4.6 Classification and prediction with longitudinal data 46
5. Discussion 50
6. Conclusions 55
7. Glossary 57
8. References 59
9. Annexes 63

Figure list

Figure 1. Gantt diagram depicting the work plan for the project. 15
Figure 2. Histograms showing the distribution of the main summary measures32
Figure 3. Network graph from pairwise Pearson correlations. 33
Figure 4. Network graph from GGM. 34
Figure 5. Component scores in the first and second component, with cases colored according to their category in the BMD variable. 37
Figure 6. Component scores in the first and fifth component, with cases colored according to their category in the fat mass distribution variable. 38
Figure 7. Component scores in the first and third component, with cases colored according to their category in the lean mass distribution variable. 38
Figure 8. Evolution in PCA scores of an example case through the years 2000 to 2015. 39
Figure 9. Variable importance plot for the complete random forest model. 40
Figure 10. Variable importance plot for the reduced random forest model. 41
Figure 11. ROC curves for 'complete' models (rf: random forest, glm: logistic regression, svmLinear: SVM with linear kernel, svmRadial: SVM with radial kernel) 45
Figure 12. ROC curves for 'reduced' models (rf: random forest, glm: logistic regression, svmLinear: SVM with linear kernel, svmRadial: SVM with radial kernel) 45
Figure 13. Distribution of FMR in the training and test sets. 48
Figure 14. Plots of predicted against observed values for FMR in the longitudinal models with a continuous outcome. 49

1. Introduction

1.1 Context and justification of the work

1.1.1 General description

Antiretroviral therapy has dramatically increased life expectancy in HIV (Human Immunodeficiency Virus)-infected patients (1). As a consequence, currently these individuals face other complications associated with HIV-infection and treatment, not directly caused by AIDS, but which have a direct and important impact over their quality of life (2). Although these complications can affect many aspects of health, there is a prominent involvement of problems in the musculoskeletal system (3,4).

A well-known complication associated to HIV is low Bone Mineral Density (BMD) (5). A significant decrease in BMD relative to the normative values in the general population may lead to a diagnosis of osteopenia or osteoporosis, which increases the risk of bone fracture and is linked to a shorter life expectancy (6). BMD is commonly measured with a dual X-ray absorptiometry scan (DXA) and compared to normative data from sex and age-matched individuals from the general population to obtain T-scores that represent the deviation from the average values (7). A T-score > -1.0 (i.e. within one standard deviation from the population mean) is considered normal. Below this value, T-scores between -1.0 and -2.49 (within 1 and 2.5 standard deviations) receive a diagnosis of osteopenia, while a T-score below -2.5 will receive a diagnosis of osteoporosis.

An additional complication accompanying HIV infection is the loss of muscle or lean mass, sometimes termed sarcopenia (3). Sarcopenia is an age-related loss of muscle mass quantity and quality (8); however, low muscle mass can also arise secondary to other conditions such as HIV. For example, Buehring et al. (3) found that between 18.8% and 21.9% of HIV-infected males presented with low muscle mass. A common criterium to consider an individual as affected by low muscle mass is the Appendicular Lean Mass Index/height^2 (cutoff points at < 7 kg/m^2 for men < 6 kg/m^2 for women) (9,10).

Alterations in fat mass distribution, generally known as lipodystrophy, are also common in HIV, although they have changed over time: as recently reviewed by Koethe et al. (2), early on in the HIV epidemic, muscle and fat loss, known as ‘wasting’, were a hallmark of AIDS. However, with the advent of combined

antiretroviral therapy (cART), a different clinical syndrome of lipoatrophy was described, primarily linked with treatment with protease inhibitors and staduvine or zidovudine. This syndrome was characterized by loss of fat mass in the face and limbs, and in some cases also by an increase of fat mass in the trunk, leading to a combination of lipoatrophy and lipohypertrophy termed 'lipodystrophy'. Currently, the use of new-generation antiretroviral drugs with less toxicity has led to a reduction in the prevalence of lipoatrophy, while lipohypertrophy has become the most common fat-related alteration in HIV-infected individuals receiving cART (2). Lipodystrophy is defined using the Fat Mass Ratio (FMR) (11), that is, the ratio between the percent of trunk fat mass and the percent of lower-limb fat mass, with cutoff values of FMR >= 1.24 for men and >= 0.95 for women to be considered as affected by lipodystrophy. However, in clinical settings diagnosis still relies on anthropometric measures, which require a loss of 30% or more of limb fat (as measured with DXA) for lipodystrophy to become clinically evident (12). This, in addition to the difficulty in discriminating lipodystrophy from generalized fat gain, may leave out subtle or early cases that will end up becoming evident with time (2).

Dual-energy X-ray absorptiometry (DXA) is usually used to measure bone mineral density (BMD). However, it can also estimate the proportion of other bodily tissues, including fat and lean mass, and as such it has been indicated to study body composition in HIV and potential adverse effects of antiretroviral therapy (13). The present work will benefit from an existing database on DXA measurements in HIV patients to study changes in body fat, lean mass and bone distribution, with a focus in lipodystrophy. Although these problems have been previously studied, there is still a lack for descriptive and predictive tools that can anticipate the patient's evolution and outcome. This work will test the ability of automated learning algorithms for the description and classification of HIV patients in terms of alterations in body fat distribution, lean mass loss and low BMD (osteopenia and osteoporosis), based on data from DXA measurements.

1.2.1 Justification

A key aspect to improve HIV patients' quality of life is the development of accurate prognostic tools to predict the evolution of the disease and its associated complications. Machine learning tools have the potential of identifying patterns and abstract relationships in data that can be used to make this kind of prediction, and anticipate patients' progression so that problems can be minimized or even prevented. However, this approach has not been systematically tested yet. This is the aspect this work aims to explore: the development of tools based on automated learning algorithms that can summarize or describe data linked to a patient's current clinical state and their progress through time, as well as to identify the most important variables in the

diagnosis of body composition alterations. Current clinical criteria to diagnose osteoporosis, low lean mass or lipodystrophy consider only a small part of the information generated by the DXA examination (e.g. a ratio of two values). Using machine learning methods, it may be possible to design new, more precise body composition indices that are based on the whole set of measurements, instead of only a few. Moreover, if we can identify the variables that are most determinant for diagnosis and/or prognosis of body composition alterations, data collection could be simplified. Finally, the application of machine learning algorithms to longitudinal data opens the door to predictions that may aid in the identification of high-risk cases, prevention, and development of quantifiable measures of aging. This work, then, involves the application of machine learning and multivariate analysis methods covered throughout the master's study plan in an integrated way to a complex dataset.

1.2 Project goals

1.2.1 General goals

• Developing descriptive tools to summarize a patient's trajectory and available information of their current status regarding DXA measures.

• Assessing different classification tools to identify the most relevant variables in the diagnosis of anomalies in body fat distribution and predict its evolution.

1.2.2 Specific goals

1.1 Summarize the information contained in the database for the total sample of patients through a reduced number of (a combination of) variables.
1.2 Generate descriptive summary measures at the individual level that describe the subject's current clinical status based on the multiple variables obtained through DXA.
1.3 Evaluate the dependencies and relationships among the variables registered through DXA examination to better understand tissue distribution.

2.1 Evaluate and compare the performance of different machine learning classifiers in the cross-sectional prediction of body fat distribution alterations based on DXA measurements to identify the variables that better characterize patients with anomalies.
2.2 Evaluate and compare the performance of linear mixed models and machine learning algorithms in the classification and prediction of body

fat distribution alterations incorporating the longitudinal structure of the dataset.

1.3 Methodological approach

The dataset to be analyzed is a complex database with multiple measurements for each case, missing data and highly correlated variables. Therefore, the first step will be a description of complete and incomplete cases, available measurements and coding errors. Once the dataset is prepared for the analyses, we will use different multivariate analysis and machine learning methods for description and prediction.

To study the relationships and dependencies among variables, and to generate summary measures, we will use multivariate dimension reduction methods. Particularly interesting for this dataset, where we expect that variables will be highly correlated (e.g. lean mass from the left arm is expected to correlate with lean mass from the right arm), is Principal Component Analysis (PCA). PCA is a multivariate technique aimed at reducing the dimensionality of a multivariate dataset, while accounting for as much of the original variation as possible, by generating new variables (the *principal components*) that are linear combinations of the original variables, uncorrelated and ordered so that the first few components account for most of the variation in the whole original dataset (14). On the other hand, Gaussian Graphical Models (15) are a novel approach to study the relationships between biological variables and infer dependencies among them by obtaining full order partial correlations (correlations between pairs of variables that control for the effect of all the other available variables). These relationships can later be represented in network graphs that visually summarize the dependency structure of a dataset. Depending on the distribution of the variables in the dataset (if data do not fit a multivariate normal distribution), Mixed Graphical Models can be used instead (16). These are known as *unsupervised* machine learning methods.

To identify the variables that better characterize those individuals with fat distribution anomalies, we will need to use different methods, particularly supervised machine learning algorithms (e.g. decision trees/random forest, logistic regression or support vector machines). To do this, we will consider different algorithms and evaluate their fit to the type of information contained in the dataset to select a set of models that will be tested and fine-tuned to obtain the best possible predictions. This will be also the procedure for longitudinal prediction. In this last case, prediction can be applied to classification or to numeric prediction (i.e. predict the future values of a specific variable). Since the dataset has a repeated measures structure, we will use both parametric models (mixed models) and data-driven (machine learning) models. To avoid

bias in model performance, data will be divided in training and test sets and use cross-validation for model selection, therefore ensuring that results are independent of data partitioning.

To develop this work, we will use the R programming language and environment. R is a free, open-source software especially fit for statistical analysis and with numerous resources for machine learning. Currently, multiple libraries and community-created content are available to perform the proposed analyses, which ensures the feasibility of the project. Moreover, it has been extensively used throughout the master. Thus, it has some advantages over other programming languages such as Matlab (which requires payment), or Python (covered in less depth in the master's study plan). Additionally, R can be combined with Markdown (using RMarkdown) to generate dynamic reports with the analysis code and ensure reproducibility of the research.

To facilitate writing the report and organizing information, Mendeley will be used as the reference manager. All software to be used in this project is free, so no economic costs will be associated with its development.

Finally, ethical aspects of the project have been considered. This work will employ a database with information from DXA examinations from human subjects acquired for different observational studies that do not involve any experimental manipulation or pharmacological intervention. The database is fully anonymized, and identification of any individual subject is not possible. Therefore, it can be used for the project without restrictions.

1.4 Work plan

1.4.1 Tasks

- Literature review. This will use the PubMed database (https://pubmed.ncbi.nlm.nih.gov/) as the primary source of information, mainly in the form of scientific papers. Reference lists in the revised papers will also be used to get other studies not found directly by search. Additional searches may use general search engines (e.g. Google) to access sources other than scientific literature. This task may be divided in the following subtasks:
    - Review of the scientific literature on bone, fat and muscle mass distribution in the context of HIV infection.
    - Review of DXA measures and their interpretation, establishing cutoffs and working definitions for lipodystrophy, low lean mass and osteoporosis.

    - Review of data analysis and machine learning methods suitable for use with DXA methods.
    - Identification of R packages necessary for application of data analysis and machine learning methods. These may be primarily accessed through the Comprehensive R Archive Network (CRAN).

- Database preparation.
    - Exploration of the database: identification of missing data or incomplete cases, duplicates and errors.
    - Identification and definition of the recorded variables.
    - Extraction of basic summary statistics and quality checks.

- Extraction of summary measures for individual clinical status.
    - Application of dimension reduction or unsupervised machine learning algorithms to obtain summary indices of patients' clinical status.
    - Application of correlational methods (e.g. GGMs or MGMs) to explore relationships and dependencies among variables.

- Training machine learning and parametric mixed models for classification and prediction. Training different models with different tuning parameters.
    - Training models for cross-sectional classification/prediction.
    - Training models for longitudinal prediction.

- Interpretation and discussion of results.

- Writing the report.

- Preparation of the presentation and defense.

Figure 1 shows a Gantt diagram depicting the work plan.

1.4.2 Calendar

Task	Start	End	Duration (days)
Literature review	21/09/2020	25/10/2020	34
Body fat, bone and muscle distribution in the context of HIV	21/09/2020	13/10/2020	22
DXA measures and cutoffs	14/10/2020	21/10/2020	7
Data analysis and machine learning methods	14/10/2020	25/10/2020	11
R packages	21/10/2020	25/10/2020	4
Database preparation	14/10/2020	25/10/2020	11
Exploration of the database	14/10/2020	25/10/2020	11
Identification of variables	14/10/2020	16/10/2020	2
Basic summary statistics and quality checks	19/10/2020	25/10/2020	6
Summary measures for clinical status	26/10/2020	16/11/2020	21
Extraction of individual summary measures	26/10/2020	06/11/2020	11
Exploration of variable dependencies	07/11/2020	16/11/2020	9
Training models for classification and prediction	17/11/2020	14/12/2020	27
Cross-sectional models	17/11/2020	29/11/2020	13
Longitudinal models	30/11/2020	21/11/2020	22
Interpretation and discussion of results	15/12/2020	30/12/2020	15
Writing the report	14/10/2020	05/01/2021	83
Preparation of the presentation and defense	01/01/2021	10/01/2021	9
Public defense	13/01/2021	20/01/2021	7

1.5 Brief summary of the resulting products

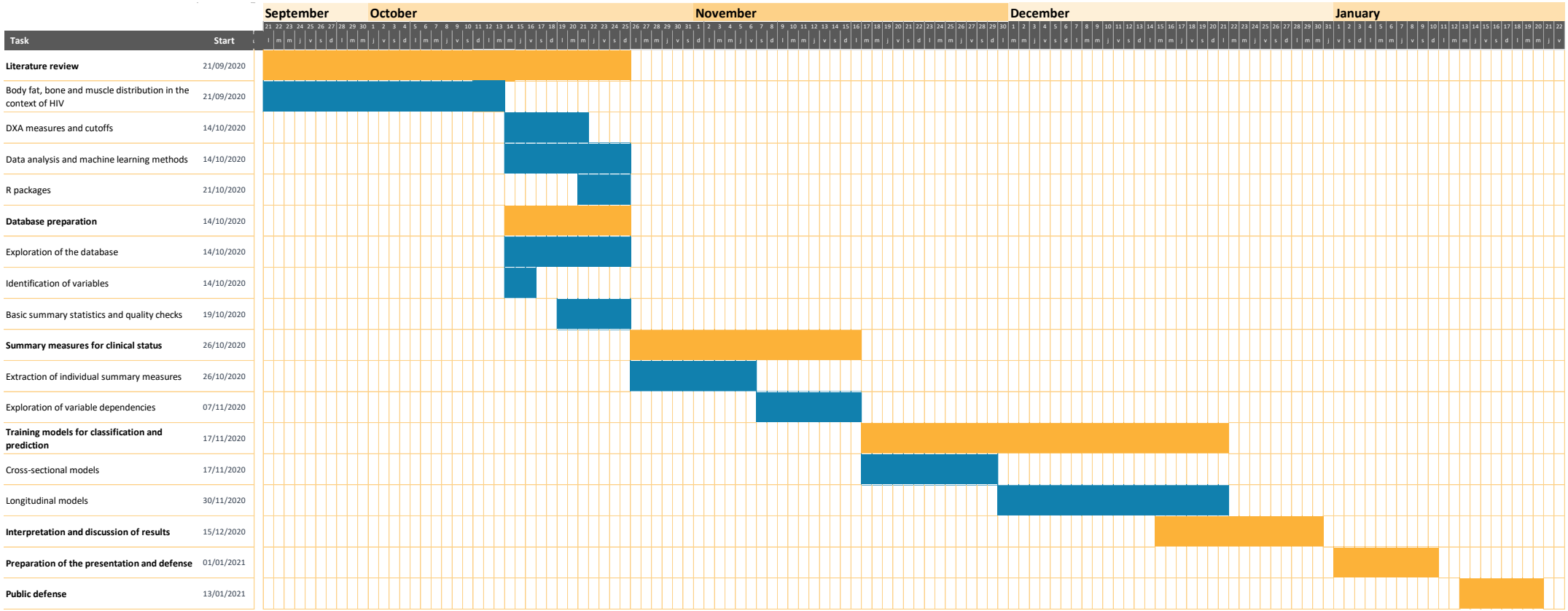
The work developed in this project has resulted in a curated dataset, obtained from the original data after cleaning and preparation for the analyses, which is attached as Annex 4 (the original dataset is attached as Annex 2). The full code for all analyses is also included in the Annexes, and covers data management and preparation to obtain the curated dataset from the original dataset, correlation analyses, PCA, training and test of machine learning models for cross-sectional classification, and training and test of mixed and machine learning models for classification and prediction of longitudinal data. The code is included as RMarkdown files and their compilated output in html format. The present report and the oral presentation are also resulting products from this project.

1.6 Brief summary of the remaining chapters

This report is divided in the following chapters:

- Background: this chapter covers two important aspects of the work. The first is a description of the problem the project is aimed to address, namely body fat distribution alterations, or lipodystrophy, in HIV infection. The second is a brief review of the methods to be used to work with the type of data included in the project, including methods to work with highly correlated data, common machine learning algorithms for cross-sectional classification, and analysis approaches for repeated measures data.
- Materials and methods: description of the data analysis methods used in the project and the steps that were followed for data management and curation, data analysis, and model construction and assessment.
- Results: detailed results obtained after the application of the methods described in the previous chapter.
- Discussion: comments on the obtained results, in the context of the topic of the project and the goals that were initially planned.
- Conclusions: final evaluation of the project goals and the degree to which they were achieved, evaluation of the adherence to the work plan, and proposals for future works.
- Glossary: definition of important terms and acronyms used in the report.
- References: scientific literature and bibliographic sources consulted.
- Annexes: supplementary material attached to the report.

Figure 1. Gantt diagram depicting the work plan for the project.



2. Background

The human immunodeficiency virus (HIV)[1] targets the immune system by destroying and impairing the function of immune cells, which makes affected individuals more vulnerable to infections and some types of cancer. If untreated, the patient can become immunodeficient, and develop acquired immunodeficiency syndrome (AIDS), the most advanced and severe stage of HIV infection. HIV is still a major global public health issue, with 38 million people living with the infection worldwide at the end of 2019 (17). Efforts to guarantee access of HIV-infected individuals to health services providing prevention strategies, diagnosis, treatment and care have resulted in HIV becoming, in most cases, a manageable chronic health condition (1). This, however, has resulted in new challenges in the management of health problems in people living with HIV.

2.1 Lipodystrophy in the context of HIV infection

Although complications faced by HIV-infected individuals can affect many aspects of health, there is a prominent involvement of alterations in the musculoskeletal system (3,4), which include low bone mineral density (BMD) leading to clinical diagnoses of osteopenia or osteoporosis, low muscle mass, and fat distribution alterations known as lipodystrophy. Fat and muscle mass loss can be observed in treatment-naïve patients with HIV infection, a condition also known as ‘wasting’, which was a hallmark of AIDS before the advent of antiretroviral therapy (2). This was a result of a profound caloric deficit due to opportunistic infection, altered gastrointestinal function and increased energy expenditure, generating a situation similar to starvation (2). With the introduction of combined antiretroviral therapy (cART), a new pattern of altered body fat distribution was observed, characterized by a loss of subcutaneous fat in the trunk, face and extremities, termed ‘lipoatrophy’, and an increase of trunk visceral fat, and also fat accumulation in the breasts and upper back (dorsocervical fat pad) which has been termed ‘lipohypertrophy’ (18). These two patterns can appear separately or in combination, and are commonly classified into a single ‘lipodystrophy’ phenotype (2).

Estimates of the prevalence of lipodystrophy in HIV-infected population vary widely between studies. For example, Buehring et al. (3) found that between 40 and 55% of patients showed lipodystrophy, and that it was more common in those receiving cART, but other reports have estimated the prevalence of body composition alterations between 11 and 83% (19). The absence of a clear-cut definition of lipodystrophy criteria has contributed to these disparate findings. Diagnosis of lipodystrophy has usually relied on clinical examination or semi-quantitative clinical scores (2), which depend on the expertise of the examiner, making it difficult to identify cases in small centers or to obtain an objective

[1] Although there are two types of HIV, here we refer to HIV-1, the most common and infective variant, which is present worldwide.

measure for research. A first objective definition for HIV infection lipodystrophy was proposed by Bonnet et al. (20): the fat mass ratio (FMR), defined as the ratio of the percentage of trunk fat to the percentage of legs fat as measured by DXA (Dual-energy X-ray absorptiometry). DXA is a medical imaging technique that uses spectral imaging to measure body composition. It is able to discriminate three different tissue types based on their specific X-ray attenuation properties: bone mineral content, lipid (fat mass) and lipid-free soft tissue (lean mass), that also includes water (21). Although it uses X-rays, the radiation dose delivered to the patient is small, making it a safe and widely used medical tool. As such, it has been recommended for the study of lipodystrophy in HIV-infection (13). Newer DXA systems are also able to differentiate subcutaneous and visceral abdominal fat, which can provide more accurate estimations of cardiovascular risk (21). The DXA-based measure of lipodystrophy seemed to allow a more accurate and early diagnosis than clinical definitions, although it was based only in male population. However, FMR tends to be significantly higher in males than females, which has led to the definition of gender-specific thresholds (11). Currently, cutoff values are set at FMR >= 1.24 for men and >= 0.95 for women to be considered as affected by lipodystrophy. Besides diagnosing lipodystrophy in the context of HIV infection, DXA is also used to diagnose other types of fat distribution alterations, like familial partial lipodystrophy, with high reliability (22).

Lipodystrophy in HIV infection has been mainly linked to long-term cART, especially with protease inhibitors (2). Early studies already showed that highly active antiretroviral therapy was associated with increased fat mass in the trunk and reduced fat mass in the limbs even in the absence of significant differences in total fat mass or weight relative to other medication regimens (4). Dubé et al. (23,24) showed that, in the first months after initiation of treatment, limb fat increases, which was explained as a ‘return to health’ phenomenon after fat loss as a consequence of HIV-infection. After this increase, there was a progressive loss in limb fat accompanied by an increase in trunk fat, especially in patients treated with didanosine and stavudine. Similarly, Gundurao et al. (25) reported an association between treatment with protease inhibitors and truncal obesity, face lipoatrophy, dyslipemia and hyperglycemia. Treatment with stavudine was linked to face lipoatrophy, hypertriglyceridemia and low HDL (high-density lipoprotein) cholesterol. In fact, lipodystrophy is not only defined by the morphological changes due to fat redistribution, but also by metabolic disturbances, including dyslipemia and insulin resistance (26). Besides antiretroviral use (especially nucleoside analogues and protease inhibitors), increasing age, markers of disease severity and duration of therapy have also been identified as risk factors for the development of lipodystrophy (26).

On the other hand, lipodystrophy can be a risk factor for other health complications. For example, in one study, HIV patients with lipodystrophy had a faster increase in waist circumference in a longitudinal follow-up, which may contribute to increased cardiovascular risk (18), since increased visceral fat is associated, both in HIV-infected and uninfected populations, with cardiovascular risk factors such as insulin resistance, low HDL cholesterol, and high triglycerides (27–30). A review by Calvo and Martinez (31) found that lipodystrophy is associated with inflammation, dyslipemia, diabetes,

hypertension and functional decline. Functional impairment is not only associated with lipodystrophy, but also with low muscle mass and low bone mineral density (31). Metabolic syndrome may also be associated to lipodystrophy, resulting in worse clinical outcomes. However, metabolic alterations in HIV-infected persons differ from those found in non-HIV persons with metabolic syndrome (31). Interestingly, a cross-sectional study showed that FMR was a better predictor for insulin resistance and glucose alterations than lipodystrophy defined on clinical criteria (32,33). Therefore, identifying risk factors and variables that characterize individuals with body tissue distribution alterations, and those that predict the development of lipodystrophy and other anomalies, may aid in the diagnosis and prevention of further health complications including metabolic syndrome and cardiovascular risks. With this aim, the present work will analyze DXA-derived measures of body composition, to study the associations between different body composition variables and their links to objective criteria for lipodystrophy, osteopenia/osteoporosis and low muscle mass. Moreover, it has been shown that the alterations in body tissue distribution are also interrelated, with the largest overlap found between low BMD and lipodystrophy, while lipodystrophy and low muscle mass tend to be negatively correlated (3,34). Thus, we will also test if measuring different tissues could provide information on the others (e.g. if lipodystrophy can be predicted by bone density or lean mass measures).

2.2 Methods to analyze highly correlated data

The output of DXA scanning is a set of measures for bone mineral density in different body parts (e.g. vertebrae and femur, with several measures in each case), lean mass, and fat mass and percentage (in arms, legs and trunk). This leads to a dataset where variables are highly correlated. For example, measures in the right arm are expected to correlate with the left arm, measures in one vertebra and expected to correlate with the adjacent one, and measures of the same tissue type are expected to correlate with each other. Moreover, if we scan the same individual in different moments, we also expect to observe a high correlation between these different measurements.

These features need to be kept in mind when analyzing the dataset resulting from DXA scans with repeated measures. For example, the presence of highly correlated predictors (multicollinearity) can greatly influence the outcome of regression analysis and limit the accuracy of inference with parametric models. Including highly correlated predictors in a regression analysis can lead to very imprecise estimates for the coefficients (high standard errors) and changes in the statistical significance of variables within the model depending on the inclusion or exclusion of specific variables, although predictions can still be accurate.

To deal with highly correlated data, a first step is the exploration of the correlation structure, to understand how variables are related to each other. This can be done with classical Pearson correlations, which can be graphically represented in a correlation matrix. However, ordinary pair-wise correlations do not distinguish direct from indirect effects, i.e. a high correlation between two

variables does not necessarily mean they are dependent on each other, since the correlation could be explained by the effect of a third variable. This is especially relevant in highly correlated datasets, where many variables show high correlation with each other. In this case, to distinguish direct from indirect effects, Gaussian Graphical Models (GGMs) can be used (15). This is a novel approach based on obtaining full order partial correlations between each pair of variables from the full set of variables under investigation, i.e. the correlation between a pair of variables controlling for the effect of all remaining variables in the dataset. This is done for each pair of variables and graphically represented in network form: each variable is a node of the network, and edges that link nodes represent the dependency between the variables, with a weight that represents the strength of the association. If the number of variables is smaller than the number of observations, GGMs can be estimated by inverting the covariance matrix. However, there is still a risk of overfitting the model. To reduce overfitting, there are different techniques of parameter regularization that improve partial estimates for small and moderate sample sizes, for example those based on the least absolute shrinkage and selection operator (LASSO), or based on covariance shrinkage, among others (15). GGMs can be implemented in R using different packages, like the *qgraph* (35), which can perform model selection using Bayesian Information Criteria (BIC) and plot the resulting GGM in network graph form.

A second, often complementary, approach once the correlation structure of the dataset has been examined, is to apply a method for dimension reduction, like principal component analysis (PCA). PCA works by generating a new set of variables, the *principal components*, that are the result of linear combinations of the original variables, and that fulfill the following requirements: (i) they are orthogonal, i.e. uncorrelated with each other, and (ii) they are ordered so that the first one accounts for most of the original variation in the dataset, the second one accounts for most of the original variation not explained by the first one, and so on. This way, a few principal components can account for most of the original variance, and the components can be used to graphically summarize data or as the inputs for multivariate analyses (14). Another interesting aspect of PCA is that one can extract the correlation between each variable and the principal components (called *loading*). Although principal components are not necessarily interpretable, examination of the component loadings may shed some light into which part of the variation is being captured by each component. In a dataset composed by the outputs of DXA scans, it would be expected that measures of the same tissue type would cluster together, providing components that can be considered summary measures of the original variables. Moreover, components may also capture patterns in the associations between variables that reflect anomalies, like lipodystrophy, usually measured using a simple ratio (FMR), but taking into account other variables that might also be related to the alteration. As an unsupervised machine learning method, PCA is suited to uncover hidden patterns in the data, which may be useful to enrich our understanding of what defines an alteration like lipodystrophy.

2.3 Machine learning methods for classification and prediction

As indicated in the Introduction, one of the project's aims is to assess the performance of different machine learning methods to classify HIV-infected patients depending on whether they show lipodystrophy based on the DXA variables. This also needs to be done keeping in mind the highly correlated nature of the dataset and the repeated measures structure. The first approach taken here is based on cross-sectional classification: taking just one measurement per subject, we can try to use DXA variables to classify individuals into two categories ('lipodystrophy' and 'normal', based on the FMR cutoff). For this, three different machine learning algorithms were chosen and two types of models were assessed with each one. The first included all the available predictor variables (gender, age, weight, height and all DXA measures). From these models, high accuracy should be expected, since they include the variables originally used to classify individuals according to the presence or absence of lipodystrophy. The second type of model will exclude fat mass and fat percentage variables, based on previous evidence that there is a considerable overlap between lipodystrophy and BMD alterations, and a negative correlation between lipodystrophy and low muscle mass (3,34). Thus, we will test if lipodystrophy can be predicted from other tissue types which, if results are positive, would allow a potential simplification of the examinations (e.g. shorter scanning times).

To perform cross-sectional classification, three different algorithms were selected: random forest, logistic regression and support vector machines (SVM). These are all widely used supervised learning algorithms with good performance in classification problems (36). Random forests are powerful classifiers based on decision trees, which are built using what is called *recursive partitioning*: cases are divided into subsets based on one feature, and then split repeatedly into smaller subsets until groups are sufficiently homogeneous. The aim at each split is to select a variable that results in two groups as different as possible from each other, and as homogeneous as possible within each subset. The idea behind random forests is to use a large number of decision trees that operate as an ensemble. Each individual tree makes a class prediction, and the one with more 'votes' is selected. The available algorithms make the trees' predictions uncorrelated (or with low correlations with each other), which ensures that the prediction by committee is more accurate than that of any individual tree. In addition, tree-based models are robust to correlated features, so they are a good way to handle highly correlated data (37).

Logistic regression is a type of generalized linear model used when the outcome variable is categorical. In a generalized linear model, we perform regression with an outcome variable that is not normally distributed (in the present case, the distribution would be a binomial), with a link function that linearly varies with the predicted values. For each case, it gives the probability of the case belonging to one of the classes ($p(X)$). Instead of using a linear function, in logistic regression the probability of the outcome at each value of the predictor is modelled via a logistic function (equation 1), which is limited to give values between zero and one (38). However, logistic regression might be

vulnerable to the presence of highly correlated data. This is not expected to affect predictions made by the model, but may affect inference on which variables are significant predictors of lipodystrophy, so the focus will be on its predictive capacity.

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad \text{(Equation 1)}$$

To our knowledge, only one previous study had aimed to classify HIV patients with and without lipodystrophy using automated learning methods, particularly logistic regression (39). The model aimed to predict clinically diagnosed lipodystrophy (i.e. diagnosis was not based on FMR) using demographic, clinical, metabolic and body composition variables. The final model included as predictors age, sex, duration of HIV infection, HIV disease clinical stage, waist to hip ratio, anion gap, HDL cholesterol, trunk to limb fat ratio, intra-abdominal to extra-abdominal fat ratio, and percentage leg fat, and achieved 79% sensitivity and 80% specificity. Interestingly, even though this model was formulated before Bonnet et al.'s (20) proposal of FMR as a cutoff criterion for lipodystrophy, the model included the trunk to limb fat ratio, which is a very similar measure. In this work we will also use logistic regression, although our dataset is limited to age, gender, height, weight and DXA outputs, but will be useful to compare performance with this earlier model and with other machine learning algorithms.

The third type of supervised learning algorithm that is used in this work for cross-sectional classification is SVM. This algorithm creates a representation of all the cases in an n-dimensional space, with n = number of features, and tries to find a surface, called a *hyperplane*, that creates a boundary between data points separating the outcome categories (36). The most basic version of SVM classifiers require data to be linearly separable, although they are easily extended to the case of non-linearly separable data by using *soft margins*, i.e. allowing some of the cases of one category to fall within the space of the other category (how much of this deviation is allowed is determined by a cost parameter, *C*). Another way of dealing with non-linear data is to use a kernel which maps the problem into a higher dimension space that can make non-linear relationships to appear linear. Essentially, this involves constructing new features that express the mathematical relationships between measured characteristics, allowing SVM to learn features that were not originally present in the data. Therefore, using SVM with DXA scan data would make it possible for the model to discover hidden relationships between the measured variables.

A prominent feature of the dataset used in this project, and which is often found when dealing with real world medical data, is the repeated measures or panel structure. Longitudinal data offer a unique opportunity to observe the change of a variable over time, but also present some drawbacks. Not only variables that measure similar features are expected to be correlated, but also different measurements of the same subjects are expected to show a high correlation (within-subject variance is expected to be much lower than between-subject variance). Therefore, statistical analysis of longitudinal data requires methods that can deal with such intra-subject correlation.

Linear mixed models are a way of analyzing repeated measures or longitudinal data. They take general (or generalized) linear models as the basis, and then add a 'random effects' matrix to the 'fixed effects' matrix that represents the independent variables. In practical terms, we end up with a matrix of predictors $\mathbf{X}$ with n rows (n = number of observations) and i columns (i = number of independent variables) to which we need to add a $\mathbf{Z}$ matrix that codes the 'subject' factor as a random effect, with n rows and j columns (j = number of subjects). Then, a mixed effects linear model would have the form shown in Equation 2. In the same way each predictor will be multiplied by its own coefficient, there will also be a coefficient for each subject (40).

$$y = X\beta + Zu + \epsilon \quad \text{(Equation 2)}$$

Mixed models assume three sources of variation for the data: 'systematic variation' refers to the variation in the response as a function of the covariates (the same variance captured by regular linear models), 'random effects' or random variation from subject to subject, and 'within-subject variation' or random variation from observation to observation in the same subject (41). Mixed models can assume different slopes and intercepts for different subjects, allowing to capture subject to subject variation in the model. As a first approach, the repeated measures data from the DXA scans dataset will be analyzed with mixed models, using a generalized linear mixed model with logistic regression to predict a binary outcome (lipodystrophy vs. normal) and a linear mixed model with a continuous outcome (FMR).

Mixed models imply a parametric approach to the analysis of longitudinal data, and as such rely on assumptions that can be difficult to verify. In contrast, data-driven methods can be applied without a priori assumptions and derive insight directly from raw data. However, machine learning methods often assume that data are independent and identically distributed, and there are very few machine learning methods specifically devised to deal with repeated measures (42). A recent approach was developed by Ngufor et al. (42) to integrate the random effects structure of mixed effects models into non-linear machine learning models able to handle longitudinal data, the mixed-effect machine learning (MEml) framework. This approach was successfully applied to predict longitudinal deterioration in glycemic control measures in adults with type 2 diabetes and other several public real and synthetic datasets, showing its ability to predict longitudinal changes in real clinical outcomes with high accuracy, outperforming machine learning methods that do not account for the repeated measures. The authors combined two regression tree methods, the generalized mixed-effects regression trees (GMERT) and the random-effects expectation maximization (RE-EM) to alternatively estimate the fixed- and random-effects components of a non-linear mixed effect model (see (42), for details). Two different tree-based models were accommodated to this MEml framework: random forests and gradient boosted machine (GBM). The authors provided an R library to run MEml models (https://github.com/nguforche/MEml), which will be used here to analyze the DXA dataset taking into account its longitudinal structure.

An interesting feature of MEml is that a model can be built with a specific ‘lag’ in a way that information from the present and past observations can be used to predict the outcome variable in future visits. Thus, it can be used to predict the value of the outcome variable one, two, three and even four visits ahead. This feature exploits the potential of using longitudinal data to predict clinical changes in advance, which could be used, for example, to identify cases at risk of developing lipodystrophy in future examinations, even if their FMR is currently below the cutoff, based on their trajectories. Our aim here is to make a first exploratory application of the MEml framework to DXA data, to assess its value in the longitudinal prediction of FMR and lipodystrophy.

In summary, the present project deals with a real dataset that contains the outputs of DXA scans undergone by VIH-infected individuals under antiretroviral therapy and medical follow-up during a period spanning 17 years, with a variable number of scans per patient and missing data. The aim is to study the structure of the dataset, applying correlation analyses and dimension reduction measures to obtain descriptive data, and use machine learning methods to predict lipodystrophy both cross-sectionally and taking into account the longitudinal structure of the dataset. The application of novel methods to these data will try to answer the question of whether lipodystrophy can be predicted using automated learning methods, which would facilitate the development of tools to aid in the identification of high-risk cases, and the prevention and treatment of body tissue distribution alterations.

3. Materials and Methods

3.1 Dataset

This work used a dataset that contains measurements of DXA scans of a sample of 1480 individuals with VIH infection (1119 males and 361 females), scanned between 1999 and 2016. The dataset contained 4319 observations, corresponding to multiple examinations of these patients, and 76 variables corresponding to the measures obtained from the DXA, including fat mass, lean mass and bone density measures. The dataset also included some composite measures derived from these, such as body mass index (BMI, weight (kg)/height^2 (m)) or fat mass ratio (FMR, trunk percent fat/legs percent fat). A complete list of the original variables and their descriptions can be found in Annex 1 (Supplementary table 1).

All the described analyses were performed in R (version 3.6.1) with RStudio (version 1.2.5033).

3.2 Data management

Data were imported into R from the original Excel database (Annex 2), discarding variables not useful for the subsequent analyses, and all pre-processing steps were performed in R. Variables with date format were converted from Excel date to R date format using the package *lubridate* (43).

Cases with the same ID and examination date were considered duplicates and were removed. If any of the entries contained missing values, the one with the fewest missing values was retained. An additional case was removed due to having many missing entries (65 of the 76 variables). After duplicate removal, 4297 observations were retained.

Descriptive statistics (mean, standard deviation, minimum, maximum, range and standard error) were extracted for additional quality checks on the data. Frequency tables were obtained from categorical variables to check for coding errors, after which two cases with an incorrectly labelled gender (lowercase ‘h’ instead of ‘H’ for male), were corrected. Histograms for all numeric variables were also inspected to detect outliers or abnormal distributions. Examination of histograms and descriptive measures identified implausible values in left arm fat percentage and fat mass in 11 cases (percentages with values above 100%), and in total fat mass in one additional case. These were considered as coding errors and substituted by missing values (NAs). After performing these corrections, the composite measures were recalculated to ensure they did not include any of the values identified as errors. This also corrected errors in the calculation of composite measures due to Excel originally treating some of the missing inputs as zeroes. Descriptive statistics and histograms were extracted for all the variables after the corrections.

The final step for data management was creating categorical variables to identify body tissue distribution alterations: lipodystrophy, low muscle mass and osteopenia/osteoporosis (see Results section). A detailed, step-by-step description of the data cleaning procedure, including all the code and outputs can be found in Annex 3. The resulting curated dataset can be found in Annex 4.

3.3 Exploration of the correlation structure

The present dataset contains measurements for fat mass, lean mass and bone mineral density of different body parts, which are expected to be highly correlated. Therefore, as a first step in the exploration of the dataset we studied the correlation structure of the variables from the DXA scan. First, the variables showing skewed distributions in the histograms (arms and legs fat mass and percent fat) were transformed using the *log10()* function to make their distribution more similar to a Gaussian distribution. This was done because the correlation methods used assume that the data follow a (multivariate) normal distribution. However, running the same models without the transformation did not significantly alter the results. Then, the correlation structure was explored, first, by building a correlation matrix using the 58 variables derived from the DXA scan with Pearson's correlations, and representing it in graph form. However, in a highly correlated dataset such as the one used in this work, dependence between two variables can be explained by the effect of a third variable. Thus, in addition to using Pearson correlations, dependencies between variables were also studied through partial correlations by using a GGM. A GGM extracts full order partial correlations for every pair of variables (i.e. the correlation between two variables controlling for the effect of every other variable). The GGM was selected using the *ggmModSelect()* function from the *qgraph* library for R using graphical lasso and Bayesian Information Criteria (35). In both cases (Pearson correlations and GGM), correlations were computed for all the cases with valid data in the two variables being correlated (i.e. we used pairwise complete observations) instead of excluding subjects with any missing input (i.e. using only complete cases), to retain as much information as possible. The complete code for the exploration of the correlation structure, including all outputs, can be found in Annex 5.

3.4 Data reduction: Principal Component Analysis

Principal Component Analysis (PCA) was used to reduce the dimensionality of the dataset. The aim of PCA is to transform the original variables to a new set of variables, the *principal components*, that are linear combinations of the original ones. These new variables are uncorrelated and ordered so that the first few account for the largest amount of variation in the original dataset (14). In this way, information contained in a large number of variables will be combined into a smaller number of variables. The function *princomp()* was used to perform PCA on the same variables included in the correlation analysis (the 58 measures extracted from DXA scans). Given that these variables are recorded in very different scales, the PCA was computed from the correlation matrix

rather than from the covariance matrix. For PCA, only complete cases were used, since the functions to compute PCA do not accept missing data. Selection of complete cases was preferred over value imputation to avoid distorting variable distribution and to work with completely observational data.

3.5 Machine learning methods for prediction of fat mass distribution alterations

The above described methods, like PCA, are considered *unsupervised learning* algorithms, since they are aimed at identifying hidden relationships between the variables. For prediction and classification purposes, *supervised learning* algorithms were used. The first step for prediction of fat mass distribution alterations was cross-sectional prediction, i.e., prediction of the presence or absence of lipodystrophy in one particular observation based on the DXA scan measures acquired in the same observation. For this, we used only one measurement per subject, since using the repeated measures could bias predictions, as measurements of the same individual across different moments are likely to be highly correlated. For individuals with more than one scan, we selected the most recent observation without missing data.

Again, given that the machine learning models to be used cannot accommodate missing data, only complete cases were selected. This resulted in the loss of 27 cases from the original 1480, which is less than 2%. The subsequent analyses were performed with 1453 complete cases.

Age, gender, height, weight and the DXA scan measures were used as predictor variables, and lipodystrophy (present/absent) was used as categorical outcome. Training and testing of the models was performed using the *caret* library for R (44,45). Three different models were assessed: random forest, logistic regression and support vector machine (SVM). All the models were tested twice: one with the complete set of predictors, including all DXA measures, and one excluding fat-mass measures, to test whether non-fat measures were able to predict fat-mass distribution alterations. A detailed description of all the procedures, including all the code and outputs, can be found in Annex 6.

3.5.1 Dataset division

The data were split into training and test sets using the *createDataPartition()* function in *caret*. Two thirds of the original dataset (974 cases) were used for the training set and the remaining third (479 cases) were used for the test set. The selected function ensured that the random division preserved the same proportions of positive and negative cases (i.e. cases with and without lipodystrophy) in both sets.

3.5.2 Model training

Parameters for model training that were common for all models were specified through the *trainControl()* function in *caret*. Model training was set up with 10-

fold cross-validation to select the optimal model from the set of tested tuning parameters. Model selection was based on the AUC-ROC (Area Under the Curve-Receiver Operating Characteristic) measure. This is a commonly used measure for model performance that is based on the trade-off between the detection of true positives and the avoidance of false positives (36). The ROC curve graphically represents the true positive rate at varying false positive thresholds. A perfect classifier would have a curve passing through the points at 100% true positive rate and 0% false positive rate. The area under the curve (AUC) is the total area under the ROC curve if the ROC diagram is treated as a two-dimensional square, and ranges from 0.5 (no predictive value) to 1.0 (perfect classifier).

The proportion of cases of each category was unbalanced: approximately 74% of cases fulfilled criteria for lipodystrophy, as opposed to 26% of cases which were considered normal. This imbalance can induce bias in the model training and favor classification in the most frequent category. To overcome this limitation, we used SMOTE (Synthetic Minority Oversampling Technique), which is an oversampling technique to create synthetic data points that balance class distribution during training. This way the model is trained as if both categories were equally frequent (46), and is implemented in *caret*.

For the random forest model, the tuning parameter for optimal model selection was *mtry*. In random forest models, only a subset of predictors is selected each time a split is done (in contrast with decision trees, which consider all predictors at each split). Typically, $m = \sqrt{p}$ predictors are chosen, where p is the total number of predictors. In our case, this would be $m = 7.68$. However, this might not be the optimal numbers, so a range of possible values were tested (between 1 and 30) to select the value that yielded the best performance in the training set cross-validation. In *caret*, the values to be tested are selected through the *mtry* parameter.

For logistic regression, given that this is a generalized linear model, no tuning parameters are available in *caret*. Therefore, this model was estimated with the default settings. However, data preprocessing was applied (centering and scaling) to be able to assess the relative contribution of the predictors in the model by comparing the coefficients for each variable. Note that this does not change the predictions, but changes the interpretation of the coefficients. If we wanted to interpret coefficients as, for example, odds ratios, we should either back-transform them or run the model with the original variables. However, since our interest is the predictive value of the model, we used the scaled predictors.

For SVM, two different kernels were tested. First, we tested an SVM model with a linear kernel, i.e. without any data transformation other than centering and scaling for preprocessing. For this model, 11 different values for the cost value or *C* parameter were tested ranging from 0.03125 to 8192 (2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{2}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}). Here, greater values force the optimization to try harder to achieve 100 percent separation between the two classes, while lower values allow for wider margins in the hyperplane. The second variant of the SVM model used a Gaussian radial basis function (RBF) kernel. Non-linear kernels

allow modelling non-linear relationships within the data by adding additional dimensions through which the data can be separated. In addition to the penalization parameter *C*, when using a radial kernel we can also tune the parameter *sigma*, which produces more 'linear' results with low values, and more flexed decision boundaries with high values. For the SVM model with radial kernel, the same 11 different values for *C* were tested as in the model with the linear kernel, combined with 11 values of *sigma* (2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{2}, 2^{3}).

3.5.3 Assessing model performance

Model performance was assessed by testing the accuracy of the predictions using a confusion matrix. Relevant performance metrics were extracted for each model: sensitivity (proportion of lipodystrophy cases correctly diagnosed as lipodystrophy), specificity (proportion of normal cases correctly diagnosed as normal), general accuracy (proportion of correct predictions out of the total number of predictions) and AUC-ROC of the selected models. The kappa statistic was also considered, since this statistic adjusts accuracy by accounting for the possibility of a correct prediction made by chance (36). This is important because the proportion of positive cases (cases with lipodystrophy) was greater than the proportion of negative (normal) cases, so a model could achieve high values of accuracy and sensitivity just by assigning cases to the most frequent class. Models were compared based on these measures.

3.6 Classification and prediction with longitudinal data

To perform classification and prediction with longitudinal data, we used mixed models and the MEml framework, using lipodystrophy and FMR as possible outcome variables. For these analyses, again only complete cases were selected. Moreover, since the interest was longitudinal prediction, only cases with repeated measures were included. Given that the variables were in very different scales, numeric variables were mean-centered and scaled (divided by their standard deviation) before running any of the models. The code and detailed outputs of these analyses can be found in Annex 7.

3.6.1 Mixed effects generalized linear model: logistic regression

A mixed effects generalized linear model was used to predict the presence or absence of lipodystrophy while taking into account the repeated measures structure of the dataset. First, the dataset was divided into training and test sets, so that the model would be estimated with the training set and its predictive ability assessed with the test set. The division between training (75% data) and test (25% of the data) was done while preserving all observations of a particular individual together, so that it was assigned to the training or the test set, but not split between the two. Moreover, since the 'lipodystrophy' class was much more frequent than the 'normal' class, a simple undersampling method was applied to avoid bias in model estimation: observations in the majority class were randomly removed until the proportions of both classes were equal. This was

done in the training set, while original proportions were kept in the test set, to assess model performance in a 'real-world' situation.

Model estimation was performed using the *glmer()* function from the *lme4* R package (47). The model aimed to predict the outcome variable (lipodystrophy vs. normal) using age, gender, height, weight and all DXA measures as fixed effects and subject ID was included as single random effects term, adding a random intercept for each subject.

3.6.2 Mixed effects linear model for prediction of a continuous outcome

This model is a multiple linear regression with the addition of a random effects term to account for subject to subject variation, with FMR as the outcome variable. Given that FMR is a continuous variable, there was no need for resampling for this model, so 67% of the observations were assigned to the training set and the remaining 33% to the test set. Again, like in the previous model, all observations belonging to the same subject were preserved together.

The model was estimated with the *lmer()* function from *lme4*, including FMR as the outcome variable; age, gender, height, weight and all DXA measures as fixed effects, and subject ID as random effects term, adding random intercepts for each subject like in the previous model.

3.6.3 Mixed effects machine learning (MEml): prediction of a binary outcome

MEml is a novel framework aimed at incorporating random effects into standard machine learning algorithms, making it possible to make longitudinal predictions by using common machine learning models (42). For prediction of a binary outcome, we used the same training/test division as in the logistic regression mixed effects model to make model comparison easier. Similarly, the outcome variable was lipodystrophy class (lipodystrophy vs. normal) and the fixed effects predictors were age, gender, height, weight and all DXA measures. Subject ID was used to indicate the repeated measures. The model also requires a variable to indicate time to chronologically order different measurements. Here, time was simply defined as the number of DXA scans undergone by the subject (1 for the first, 2 for the second, and so on).

The learning algorithm used was the Mixed Effects gradient boosting machine (MEgbm), which incorporates mixed effects in the tree-based GBM algorithm (42). Model selection was performed using 3-fold cross-validation with 100 trees for the GBM, following the example provided by the authors of the method (https://github.com/nguforche/MEml/blob/master/demo/MEml_example.R).

3.6.4 Mixed effects machine learning (MEml): prediction of a continuous outcome

For prediction of FMR as a continuous outcome within the MEml framework, we used the same training/test division as in the linear mixed model above and, similarly, age, gender, height, weight and all DXA measures as predictors and subject ID to indicate repeated measures. Time was coded in the same way as

in the MEml model with a binary outcome. Two models were tested which only differed in the learning algorithm: the first used the MEgbm described above, and the second used the Mixed Effects random forest algorithm which incorporates the mixed effects structure in a random forest learner (42). In both cases, model selection was performed with 3-fold cross-validation, with 100 trees for the GBM and 400 trees for the random forest.

3.6.5 Lagged models

A novel feature of the MEml method is to make a lagged prediction, i.e. use the information from the current and past visits to predict the outcome variable in a future visit. For this, the *LongiLagSplit()* function was used to perform the separation between training and test sets. This function automatically splits the data and at the same time introduces the lagged structure. A lag = 1 was specified, so that information from one visit was used to predict the outcome variable in the next visit. This was chosen given that most individuals with repeated measures had two scans.

Two models were built, one with a binary outcome (lipodystrophy vs. normal) and one with a continuous outcome (FMR). The predictors and random effects variables were the same as in the previous MEml models. The MEgbm algorithm was again used for model construction, with 3-fold cross-validation and 100 trees.

3.6.6 Model comparison

Sensitivity, specificity, and AUC-ROC were used to compare the models with binary outcomes. Mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) were used to compare performance of the models with continuous outcomes.

4. Results

4.1 Dataset description

After removal of duplicates and coding errors, the dataset contained 4297 observations, from which 4100 were complete (i.e. contained no missing values in the DXA examination measures). These observations corresponded to 1480 individual cases (1119 males and 361 females), from which 617 had only one DXA scan and 863 had at least two scans. The time lapse between the first and the last scan ranged from 0 to 16 years, with an average of 6.08 years (SD = 4.57). Descriptive statistics for the main demographic and summary measures can be found in Table 1, and histograms are shown in Figure 2. Descriptive statistics for the complete set of variables, including separate descriptions for the first and last measurements of each participant, can be found in Annex 3.

Table 1. Descriptive statistics for the main demographic and summary measures.

	N valid	Mean	SD	Min	Max
Age	4297	45.23	9.91	18.00	82.00
Height (m)	4295	1.70	0.09	1.40	1.94
Weight (kg)	4296	68.60	12.15	32.20	120.50
Total BMD	4177	1.15	0.10	0.67	1.90
Total fat (g)	4296	15356.18	7668.50	1746.00	59508.00
Total fat (%)	4293	22.22	9.35	3.80	56.60
Total lean mass (g)	4297	50534.68	9652.71	25056.00	88914.00

BMD: Bone mineral density

4.2 Identification of body tissue distribution alterations

Body tissue distribution alterations were identified according to clinical criteria and were coded into categorical variables indicating the presence or absence of alteration.

4.2.1 Lipodystrophy

Lipodystrophy was defined through the Fat Mass Ratio (FMR), which results from the ratio between the percent of fat mass in the trunk and the percent of fat mass in the lower limbs. Cutoffs to consider a case as affected by lipodystrophy were set at FMR >= 1.24 for men and FMR >= 0.95 for women (12). From 4296 observations, 3186 fulfilled criteria for lipodystrophy (2485 male and 701 female) and 1110 (804 male and 306 female) were considered within the normal range.

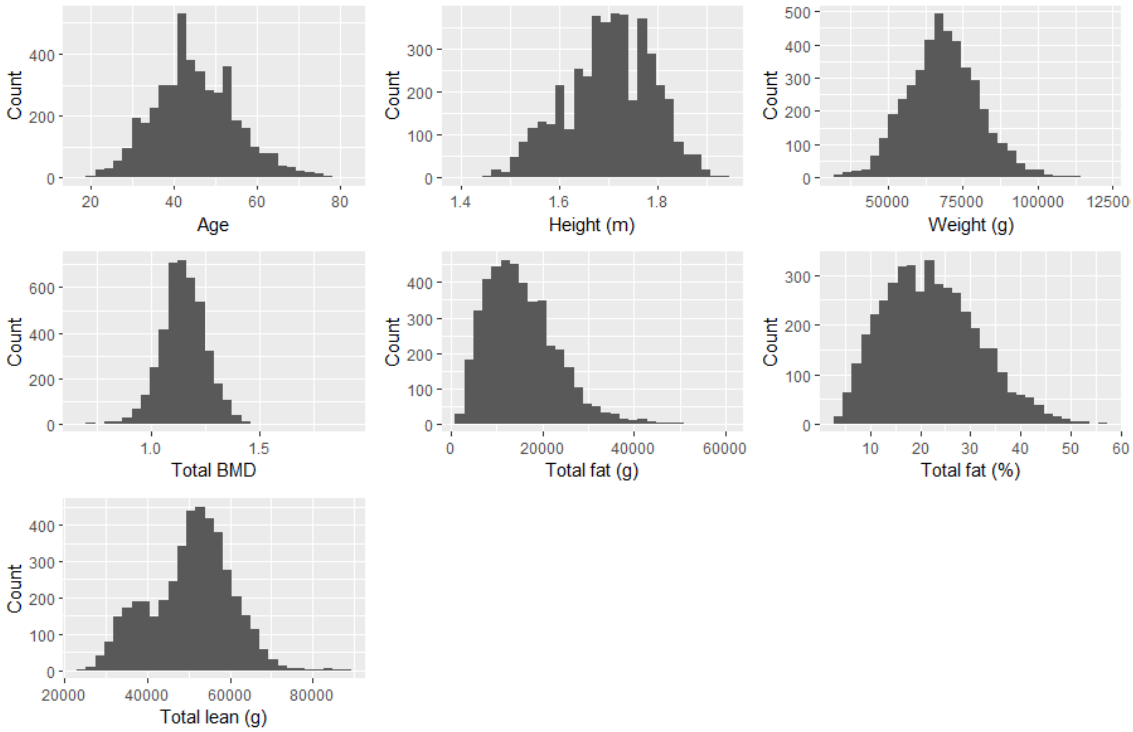


Figure 2. Histograms showing the distribution of the main summary measures

4.2.2 Low muscle mass

Low muscle mass was defined through the Appendicular Lean Mass Index/height^2, with cutoff points at 7kg/m^2 in men and 6kg/m^2 in women. Appendicular lean mass was calculated as the sum of lean mass measurements for both arms and legs. This identified 1015 observations with low muscle mass (454 male and 561 female) and 3280 (2835 male and 445 female) within the normal range (from a total of 4295 observations with valid data).

4.2.3 Low BMD: osteopenia and osteoporosis

To classify cases in terms of BMD, we used the minimum T-score from all measured body parts. With T >= -1.0, the observation was considered as normal. With T < -1.0 and T >= -2.49, the observation was classified as osteopenia, and with T < -2.49, as osteoporosis. This yielded, from 4273 cases, 851 observations with normal BMD (572 male and 279 female), 1961 with osteopenia (1537 male and 424 female), and 1461 with osteoporosis (1160 male and 301 female).

4.3 Correlation structure

The first step for exploration of the correlation structure of the dataset was to represent the correlation matrix (obtained from pairwise Pearson correlations) in graph form, as shown in Figure 3. This showed that, in general, the present dataset is highly correlated, and the variables appear to be grouped into three

clusters: measures of bone mineral density, measures of lean mass, and measures of fat mass. In all cases, variables within a cluster showed strong positive correlations among them. Moreover, bone mineral density measures correlated positively with lean mass measures, while fat mass measures correlated negatively with lean mass measures. Weight appeared grouped together with lean mass measures, although also correlated positively with fat mass, height was mainly correlated with lean mass measures, while age was shown separated from all other measures and showed negative correlations with bone mineral density and tended to show positive correlations with fat mass measures.

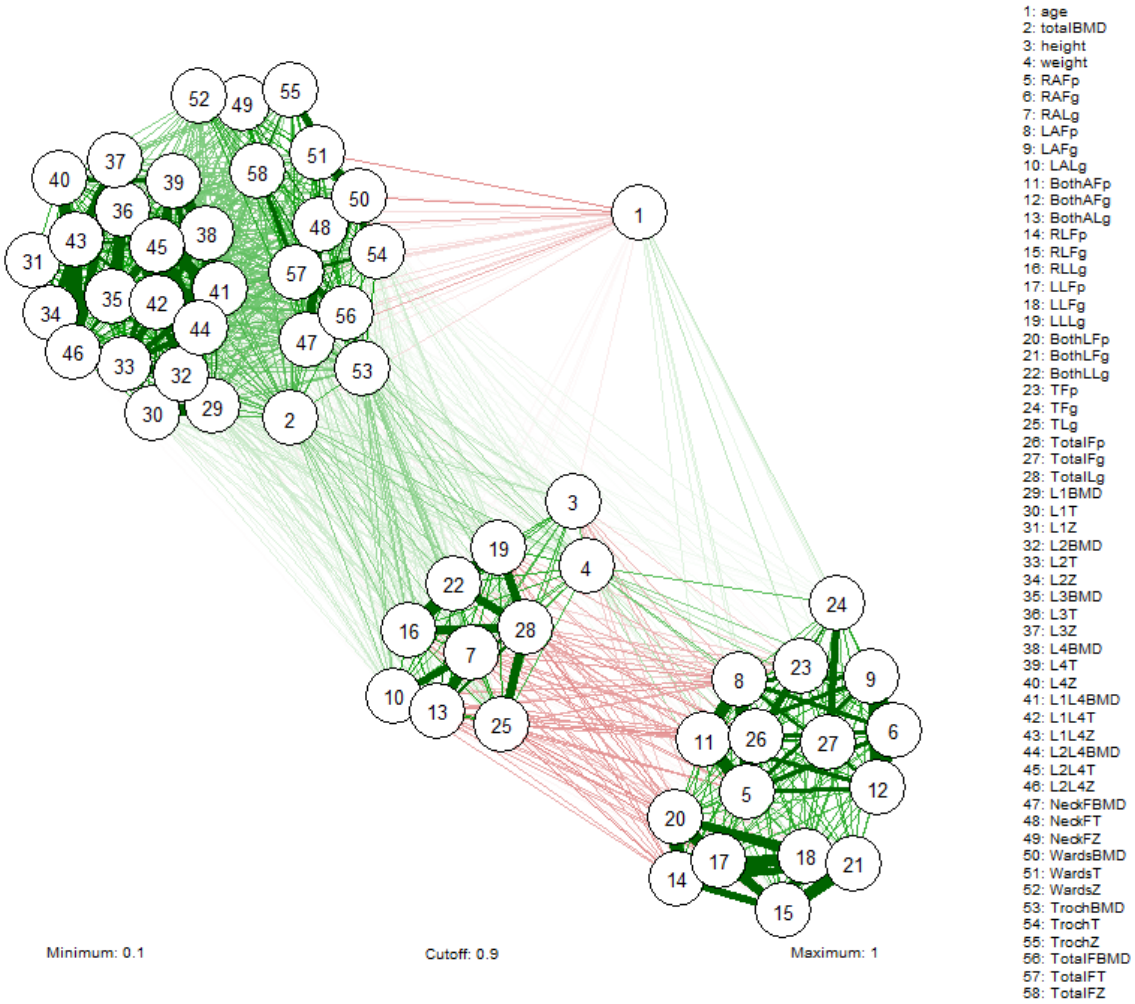


Figure 3. Network graph from pairwise Pearson correlations.

The correlation structure was further studied by means of a GGM, to obtain dependencies between variables that were significant after controlling for the effect of the remaining variables. The resulting model, shown in Figure 4, again grouped the variables in different clusters, similar to the dependency structure obtained with Pearson correlations: one corresponded to bone mineral density measures in the lumbar vertebrae; a second grouped other bone mineral density measures and also included age; a third grouped lean and fat mass

from the arms; a fourth contained lean and fat mass measures from the trunk, and a fifth contained lean and fat mass measures from the legs.

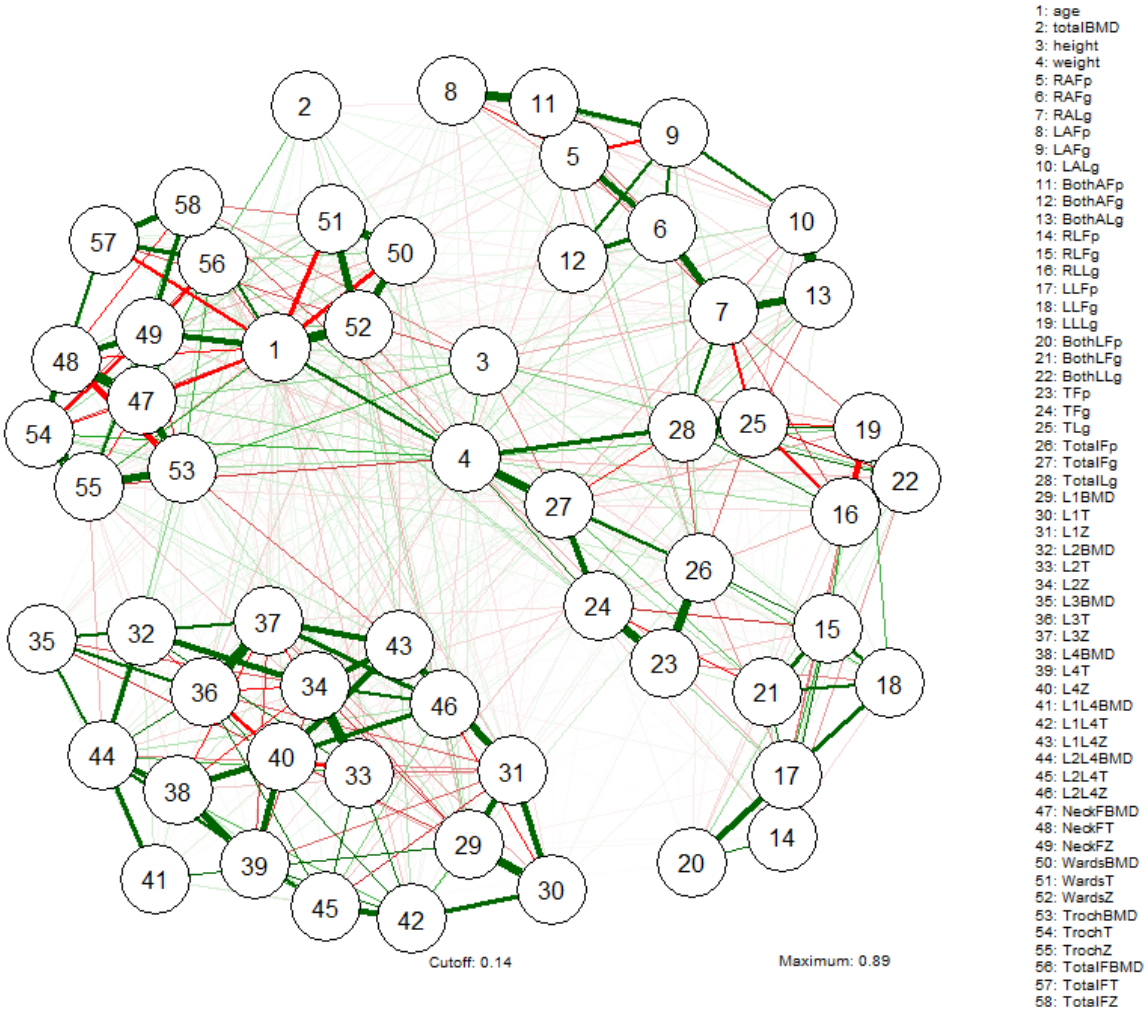


Figure 4. Network graph from GGM.

4.4 Principal component analysis

Since the function to compute the PCA does not accept missing values, only complete cases were used for the calculations, which amounted to a total of 4100 observations. Table 2 shows the results for the first ten components of the PCA. The results showed that 6 components with eigenvalues > 1 could explain more than 90% of the variance in the initial 58 variables.

Table 2. PCA results for the first ten components

	PC1	PC2	PC3	PC4	PC5
SD	4.70	3.80	2.99	2.14	1.44
Var. proprtion	0.38	0.25	0.15	0.08	0.04
Cum. proportion	0.38	0.63	0.78	0.86	0.90
	PC6	**PC7**	**PC8**	**PC9**	**PC10**
SD	1.06	0.89	0.81	0.74	0.64
Var. proprtion	0.02	0.01	0.01	0.01	0.01
Cum. proportion	0.92	0.93	0.94	0.95	0.96

4.4.1 Component loadings

Examination of the component loadings allows for an 'interpretation' of what kind of information the component conveys, although this is not always possible. Table 3 shows the component loadings for all the variables in the first six components. The first component might represent a measure of bone mineral density (with negative loadings from all bone mineral density measures). The second component has positive loadings from fat mass measures and negative, although smaller, loadings on lean mass measures. The third includes lean and fat mass measures plus some bone mineral density variables in the opposite direction. The fourth includes bone mineral density in the femur (negative loadings) and vertebrae (positive but smaller loadings). Finally, the fifth component includes negative loadings from fat mass measures in the legs and positive loadings of fat mass measures in the trunk, so it could be linked to lipodystrophy. The sixth shows very similar loadings, but in the opposite direction. Interestingly, age shows high loadings in component 5 and 6.

Table 3. Component loadings for the first six components. Loadings > 0.1 have been highlighted in bold.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
age	0.0450	0.0375	0.0098	0.0638	**0.5226**	**0.4166**
totalBMD	**-0.1692**	-0.0139	**-0.1037**	-0.0190	-0.0029	-0.0110
height	-0.0408	**-0.1075**	**-0.2052**	**0.1113**	**-0.1394**	0.0584
weight	-0.0571	0.0194	**-0.3059**	0.0963	0.0684	0.0123
RAFp	0.0141	**0.2455**	-0.0394	0.0276	**0.1175**	**-0.1996**
RAFg	-0.0085	**0.2186**	**-0.1409**	0.0558	**0.1326**	**-0.1847**
RALg	-0.0607	**-0.1296**	**-0.2440**	0.0650	-0.0169	0.0929
LAFp	0.0128	**0.2442**	-0.0377	0.0277	**0.1159**	**-0.2090**
LAFg	-0.0092	**0.2181**	**-0.1376**	0.0536	**0.1310**	**-0.1962**
LALg	-0.0605	**-0.1299**	**-0.2462**	0.0642	-0.0213	0.0876
BothAFp	0.0138	**0.2453**	-0.0386	0.0272	**0.1166**	**-0.2046**
BothAFg	-0.0087	**0.2175**	**-0.1387**	0.0564	**0.1313**	**-0.1939**
BothALg	-0.0618	**-0.1314**	**-0.2479**	0.0650	-0.0209	0.0885
RLFp	0.0183	**0.2389**	-0.0348	-0.0259	**-0.2210**	**0.1839**

RLFg	0.0005	**0.2196**	**-0.1178**	-0.0068	**-0.2371**	**0.2014**
RLLg	-0.0606	**-0.1262**	**-0.2608**	0.0696	-0.0071	0.0158
LLFp	0.0188	**0.2389**	-0.0346	-0.0268	**-0.2196**	**0.1905**
LLFg	0.0003	**0.2196**	**-0.1176**	-0.0082	**-0.2353**	**0.2069**
LLLg	-0.0608	**-0.1239**	**-0.2630**	0.0696	-0.0033	0.0150
BothLFp	0.0177	**0.2384**	-0.0348	-0.0264	**-0.2217**	**0.1864**
BothLFg	0.0008	**0.2193**	**-0.1180**	-0.0079	**-0.2371**	**0.2042**
BothLLg	-0.0609	**-0.1258**	**-0.2627**	0.0700	-0.0062	0.0168
TFp	0.0133	**0.2267**	-0.0765	0.0305	**0.1656**	0.0110
TFg	-0.0070	**0.1851**	**-0.1664**	0.0672	**0.2021**	-0.0077
TLg	-0.0560	**-0.1421**	**-0.2235**	0.0861	0.0780	-0.0717
TotalFp	0.0182	**0.2503**	-0.0516	0.0094	0.0563	0.0258
TotalFg	-0.0060	**0.2138**	**-0.1577**	0.0474	0.0730	0.0172
TotalLg	-0.0614	**-0.1438**	**-0.2503**	0.0813	0.0308	-0.0142
L1BMD	**-0.1856**	0.0097	-0.0035	**0.1250**	0.0009	-0.0857
L1T	**-0.1857**	0.0215	0.0128	**0.1173**	-0.0036	-0.0892
L1Z	**-0.1667**	0.0248	**0.1156**	**0.1043**	0.0809	0.0041
L2BMD	**-0.1913**	0.0051	0.0210	**0.1343**	-0.0399	-0.0881
L2T	**-0.1906**	0.0202	0.0408	**0.1270**	-0.0430	-0.0940
L2Z	**-0.1728**	0.0217	**0.1355**	**0.1153**	0.0291	-0.0038
L3BMD	**-0.1917**	0.0154	0.0521	**0.1272**	-0.0532	-0.0097
L3T	**-0.1913**	0.0294	0.0708	**0.1184**	-0.0550	-0.0144
L3Z	**-0.1685**	0.0306	**0.1606**	**0.1032**	0.0136	0.0700
L4BMD	**-0.1876**	0.0181	0.0337	**0.1317**	-0.0268	0.0865
L4T	**-0.1856**	0.0324	0.0509	**0.1223**	-0.0255	0.0789
L4Z	**-0.1635**	0.0337	**0.1422**	**0.1079**	0.0421	**0.1665**
L1L4BMD	**-0.1989**	0.0127	0.0285	**0.1367**	-0.0337	-0.0158
L1L4T	**-0.1994**	0.0293	0.0498	**0.1278**	-0.0357	-0.0211
L1L4Z	**-0.1779**	0.0308	**0.1495**	**0.1129**	0.0415	0.0698
L2L4BMD	**-0.1994**	0.0135	0.0371	**0.1362**	-0.0427	0.0039
L2L4T	**-0.1974**	0.0281	0.0566	**0.1272**	-0.0439	-0.0021
L2L4Z	**-0.1752**	0.0304	**0.1525**	**0.1132**	0.0304	0.0865
NeckFBMD	**-0.1669**	-0.0078	-0.0841	**-0.2189**	-0.0896	**-0.1104**
NeckFT	**-0.1653**	0.0273	-0.0395	**-0.2404**	-0.0907	**-0.1248**
NeckFZ	**-0.1553**	0.0149	0.0131	**-0.2573**	0.0831	0.0373
WardsBMD	**-0.1631**	-0.0075	-0.0387	**-0.2288**	**-0.1570**	**-0.2080**
WardsT	**-0.1615**	0.0116	-0.0150	**-0.2376**	**-0.1578**	**-0.2169**
WardsZ	**-0.1534**	0.0093	0.0368	**-0.2615**	0.0482	-0.0372
TrochBMD	**-0.1575**	-0.0220	**-0.1202**	**-0.1816**	**0.1149**	**0.1555**
TrochT	**-0.1583**	0.0404	-0.0385	**-0.2275**	**0.1132**	**0.1325**
TrochZ	**-0.1466**	0.0226	0.0198	**-0.2456**	**0.2072**	**0.2357**
TotalFBMD	**-0.1738**	-0.0076	-0.0943	**-0.2027**	0.0456	0.0039
TotalFT	**-0.1749**	0.0278	-0.0474	**-0.2269**	0.0441	-0.0075
TotalFZ	**-0.1651**	0.0092	-0.0013	**-0.2381**	**0.1846**	**0.1265**

4.4.2 Component scores

Component scores for the first six components were extracted for each observation. These scores can be potentially used to summarize a patient's status or trajectory in a few measures instead of using all the measures derived from the DXA. Figures 5-7 illustrate this by plotting the component scores according to the categorical measures on BMD (Normal, Osteopenia, Osteoporosis), lipodystrophy (Present, Absent), and low muscle mass (Present, Absent).

For BMD, given that the first component seemed to capture these measures, cases were plotted according to their scores in the first and second component and colored according to their category in the BMD variable. Figure 5 shows how cases are distributed along the first component according to their status in terms of bone mineral density, indicating that higher values in this component are linked to osteoporosis.

Component 5 grouped variables linked to leg and trunk fat in different directions, and thus may be linked to lipodystrophy. Plotting individuals according to this component, colored by the presence or absence of lipodystrophy, shows their distribution along this component. However, instead of a categorical variable this component might reflect a continuous measure of fat mass distribution alterations (see Figure 6). Since it includes variables other than the ones used to diagnose lipodystrophy, this information could be used to enrich lipodystrophy diagnosis.

Finally, the third component included lean (and, to a lesser extent, fat) mass measures, so it might be linked to low muscle mass. As shown by Figure 7, this component seems to capture low muscle mass alterations.

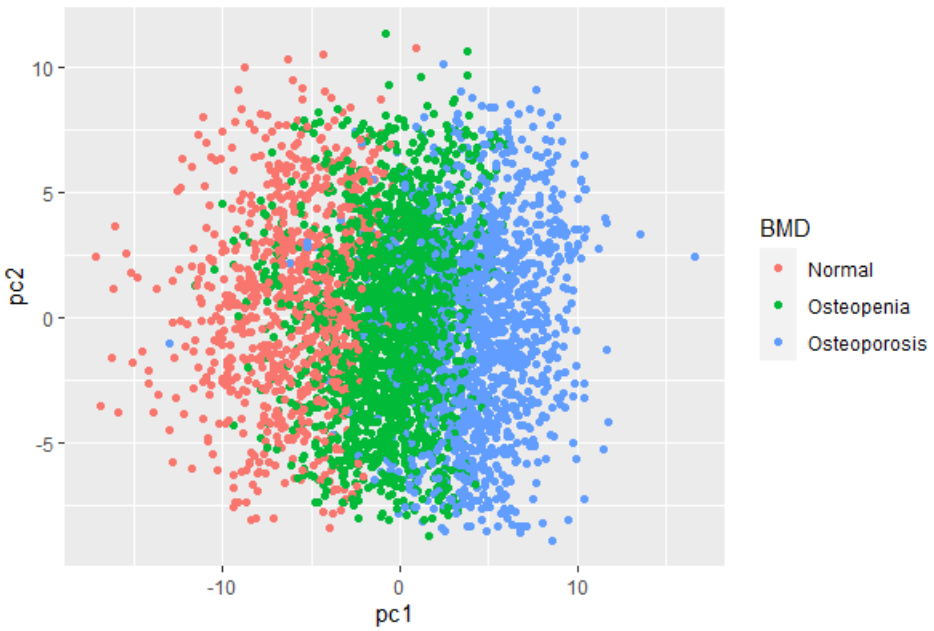


Figure 5. Component scores in the first and second component, with cases colored according to their category in the BMD variable.

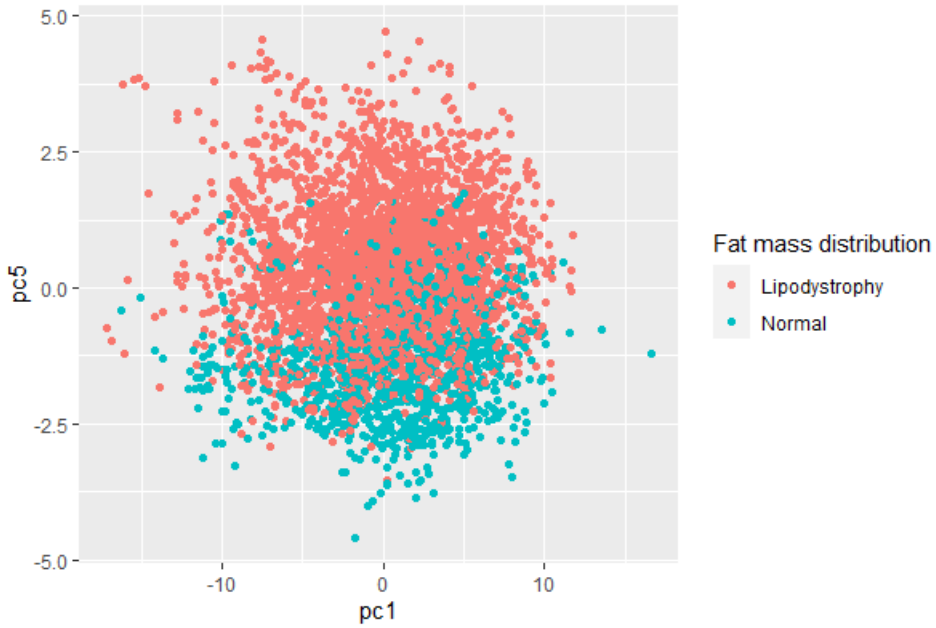


Figure 6. Component scores in the first and fifth component, with cases colored according to their category in the fat mass distribution variable.

Figure 7. Component scores in the first and third component, with cases colored according to their category in the lean mass distribution variable.

4.4.3 Example application

If these PCA scores are understood as summary measures of the clinical status of each patient, it is possible to use them to study the trajectory of a patient in each of them throughout their different measurements. As an example, Figure 8 shows the evolution of one of the patients with repeated measures from the years 2000 to 2015 in each component score. It can be observed, for example, a slight but steady increase in the scores in PC1, indicating the loss of bone

mineral density throughout the years, and an increase in PC5 and PC6, going towards greater alteration in fat mass distribution.

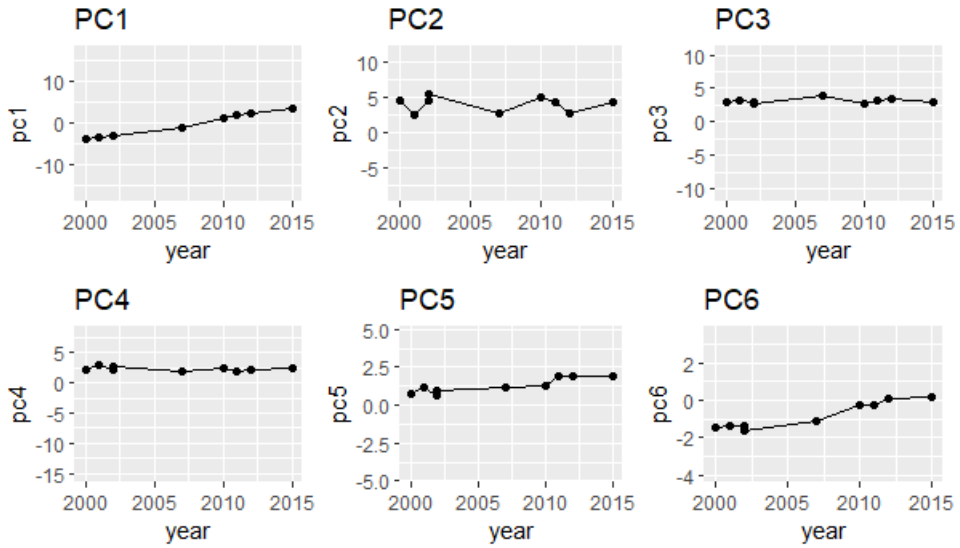


Figure 8. Evolution in PCA scores of an example case through the years 2000 to 2015.

4.5 Cross-sectional classification of fat mass distribution alterations

We used three different machine learning models (random forest, logistic regression and SVM) to classify patients into lipodystrophy and normal groups cross-sectionally (i.e. using only one observation per patient). Before model training, we checked for the presence of near-zero variance predictors, but no predictors had excessively low variance, so all of them were retained for the models. All models were tested with the complete set of predictors and excluding fat mass predictors.

4.5.1 Random forest

Using the complete set of predictors ('complete' model), the best-performing random forest model used parameter *m* = 12, which showed an AUC-ROC = 0.97, sensitivity = 0.93 and specificity = 0.88 with the training set. Performance on the test set achieved an overall accuracy = 0.92 (sensitivity = 0.93, specificity = 0.91). Figure 9 shows the variable importance plot, which identifies the relative contribution of each variable to the classification. As expected, variables quantifying trunk and leg fat were the most relevant for classification, since these are the variables used to calculate FMR and therefore to diagnose lipodystrophy.

The model that excluded fat mass variables ('reduced' model) achieved the best AUC-ROC value (0.71) with parameter *m* = 14, that showed sensitivity = 0.79 and specificity = 0.50. Performance on the test set achieved an overall accuracy = 0.74 (sensitivity = 0.81, specificity = 0.52). Figure 10 shows the variable importance plot for this model, with age as the most important predictor for lipodystrophy if fat mass measures are not available.

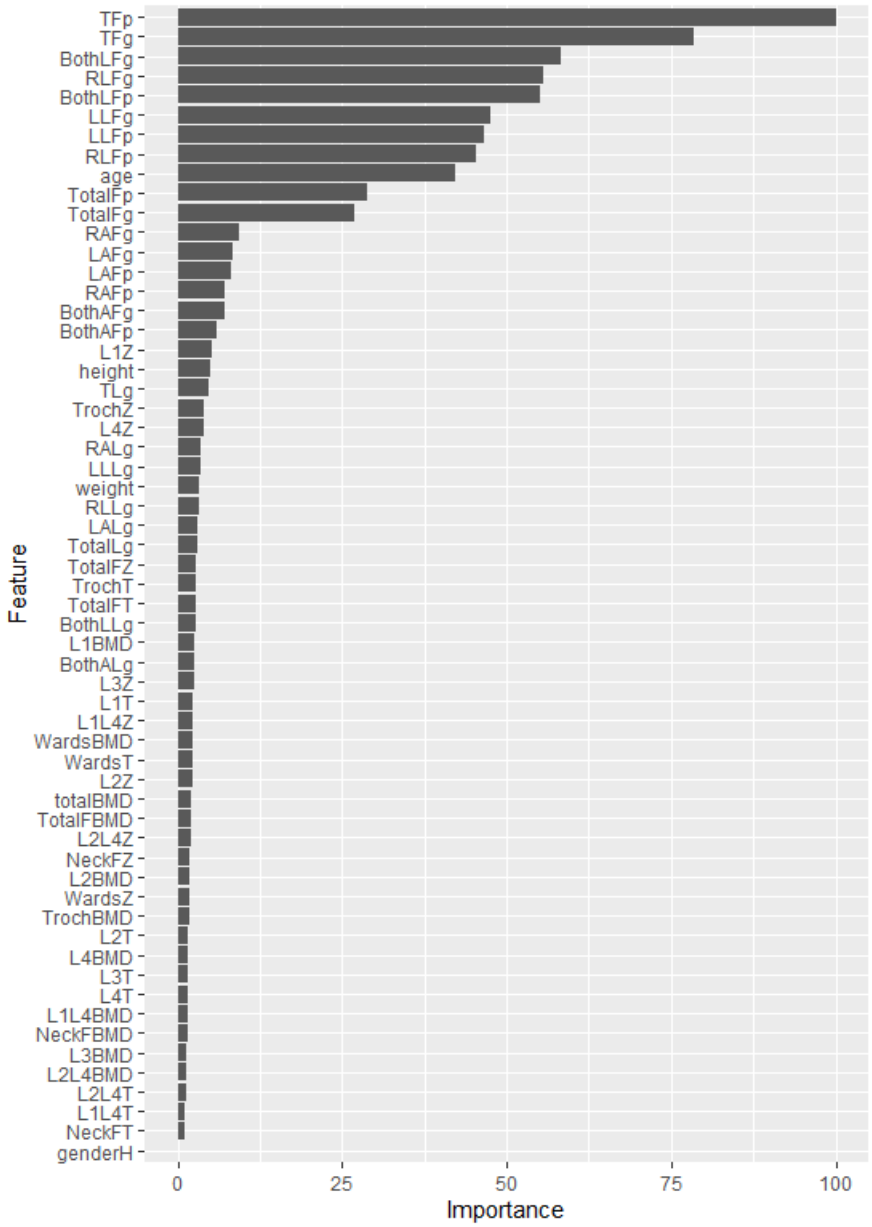


Figure 9. Variable importance plot for the complete random forest model.

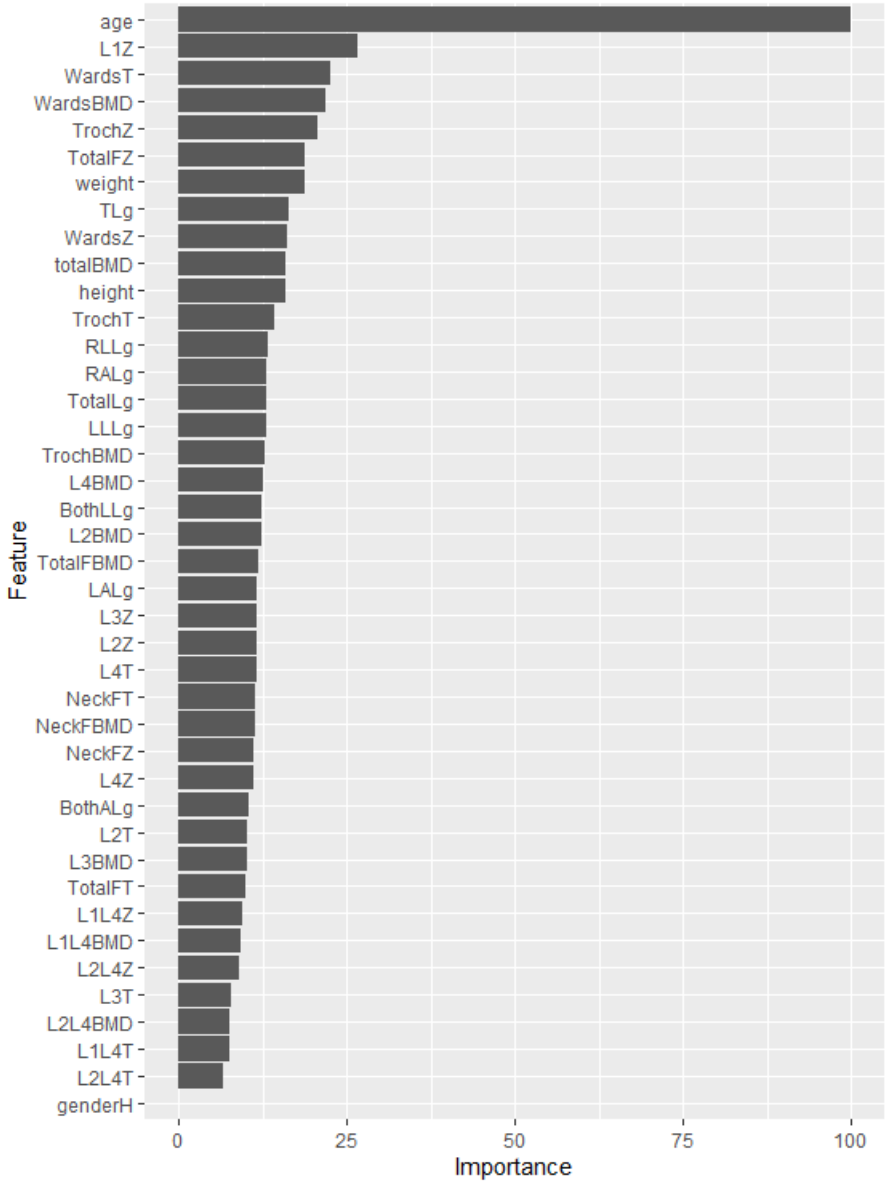


Figure 10. Variable importance plot for the reduced random forest model.

4.5.2 Logistic regression

The complete logistic regression model showed a training performance with AUC-ROC = 0.92, sensitivity = 0.94 and specificity = 0.90. Predictions in the test set achieved an overall accuracy = 0.94, sensitivity = 0.95 and specificity = 0.91. To assess variable contribution to the model, we can look at the absolute value of the coefficients for each variable. Table 4 shows the 10 largest coefficients in the model. The most relevant variables were again linked to trunk fat and leg fat, but also total fat and some bone mineral density measures.

The reduced logistic regression model showed a training performance with AUC-ROC = 0.73, sensitivity = 0.76 and specificity = 0.57. Predictions in the test set achieved an overall accuracy = 0.74, sensitivity = 0.79 and specificity =

0.60. Table 5 shows the ten largest coefficients in this model. In this case, bone mineral density measures were the most relevant.

Table 4. Ten largest coefficients from the complete logistic regression model.

Variable	Coefficient
TFp	-5.18E+15
TotalFp	3.98E+15
LLFp	2.96E+15
L3Z	-2.56E+15
LLFg	-1.89E+15
BothALg	-1.80E+15
L4Z	-1.79E+15
NeckFT	-1.68E+15
L2L4T	-1.66E+15
L1BMD	-1.65E+15

Table 5. Ten largest coefficients from the reduced logistic regression model.

Variable	Coefficient
L2BMD	-8.8291347
L3Z	-8.3193793
L3BMD	6.5815656
L2Z	5.3797179
NeckFT	-4.6693411
TrochZ	2.8613087
L2T	2.7614409
NeckFZ	2.64867
TotalFZ	-2.5771134
TotalLg	-2.5186918

4.5.3 Support Vector Machine

With the linear kernel, the complete SVM model with best performance (AUC-ROC = 0.99) was that with cost parameter *C* = 4, with sensitivity = 0.97 and specificity = 0.96 in the training set. Performance in the test set achieved an overall accuracy = 0.96, sensitivity = 0.97, specificity = 0.93. We extracted the weights of the variables in the model to assess their relative contribution. Table 6 shows the ten variables with largest weights. Like in the other models, trunk and leg fat are the variables with the highest weights.

The reduced model showed the best performance with cost parameter *C* = 0.03 (AUC-ROC = 0.74, sensitivity = 0.76, specificity = 0.56). Performance in the test set achieved an overall accuracy = 0.75, sensitivity = 0.81, specificity = 0.58. Examination of variable weights (see Table 7) shows that, like with the random forest model, after removing fat mass variables the most important predictor seems to be age, and we also find some bone mineral density and lean mass measures.

Table 6. Ten largest coefficients from the complete SVM model with linear kernel.

Variable	Coefficient
TFp	-8.8004223
BothLFp	5.2148985
LLFp	5.1772749
LLFg	-2.8973125
RLFp	2.8796285
RAFg	2.7878838
L2BMD	-2.7467311
genderH	2.7382602
L2Z	2.5331828
RAFp	-2.5164289

Table 7. Ten largest coefficients from the reduced SVM model with linear kernel.

Variable	Coefficient
age	-0.8515579
L1Z	-0.5051958
height	0.3367351
BothLLg	-0.3234204
L1L4T	-0.2940324
LLLg	-0.2722434
TotalFBMD	-0.2672395
WardsBMD	0.2613074
RALg	0.2411348
WardsT	0.2391874

With the radial kernel, the complete SVM model with best performance (AUC-ROC = 0.98) was that with cost parameter *C* = 3.59 and *sigma* = 0.01, with sensitivity = 0.96 and specificity = 0.88 in the training set. Performance in the test set achieved an overall accuracy = 0.94, sensitivity = 0.96, specificity = 0.88. Table 8 shows the ten variables with largest weights to assess their relative contribution to the model, with trunk and leg fat being the ones with larger weights.

The reduced model showed the best performance with cost parameter *C* = 4 and *sigma* = 0.0005 (AUC-ROC = 0.75, sensitivity = 0.76, specificity = 0.59). Performance in the test set achieved an overall accuracy = 0.75, sensitivity = 0.80, specificity = 0.61. Variable weights for this model are shown in Table 9. Similar to the previous models, age was the variable with the greatest weight when fat mass measures were removed.

Table 8. Ten largest coefficients from the complete SVM model with radial kernel.

Variable	Coefficient
TFp	-126.62979
TFg	-107.309099
BothLFp	106.663676
RLFp	100.545764
LLFp	98.6681925
RLFg	93.2958607
BothLFg	92.5242356
LLFg	86.9079969
genderH	41.7970114
BothLLg	-28.2151436

Table 9. Ten largest coefficients from the reduced SVM model with radial kernel.

Variable	Coefficient
age	-571.964126
WardsT	215.99297
height	212.855247
L1Z	-205.718466
WardsBMD	200.950739
TotalFZ	-160.244479
weight	-129.195914
LLLg	-114.717088
BothLLg	-93.8152496
NeckFBMD	79.8697777

4.5.4 Model comparison

Using the complete set of predictors, all models achieved very high performance, with overall accuracy values above 0.90 and AUC-ROC values close to 1, as shown in Figure 11, especially in the case of SVM models. This is unsurprising, since lipodystrophy diagnosis is established by a ratio between trunk fat and leg fat, and both these measures were included as predictors. Therefore, the models correctly captured this relationship and classified the individuals. The variable importance plots and variable weights also confirm that trunk fat and leg fat were the most relevant features for classification.

Importantly, model performance fell when using the reduced models, i.e. when excluding fat mass measures. The models showed acceptable overall accuracy and AUC-ROC values, but a close examination showed that while sensitivity was within acceptable levels, the reduced models showed poor specificity: the best performing model, SVM with linear kernel, only reached 0.60. Therefore, classification based on variables other than fat-mass measures may incorrectly lead to a diagnosis of lipodystrophy in non-affected individuals. On the other hand, as shown in Figure 12, all these models performed very similarly so they are likely reflecting the predictive ability of this set of variables independently of

the model used for prediction. Other complementary performance measures are available in Annex 6.

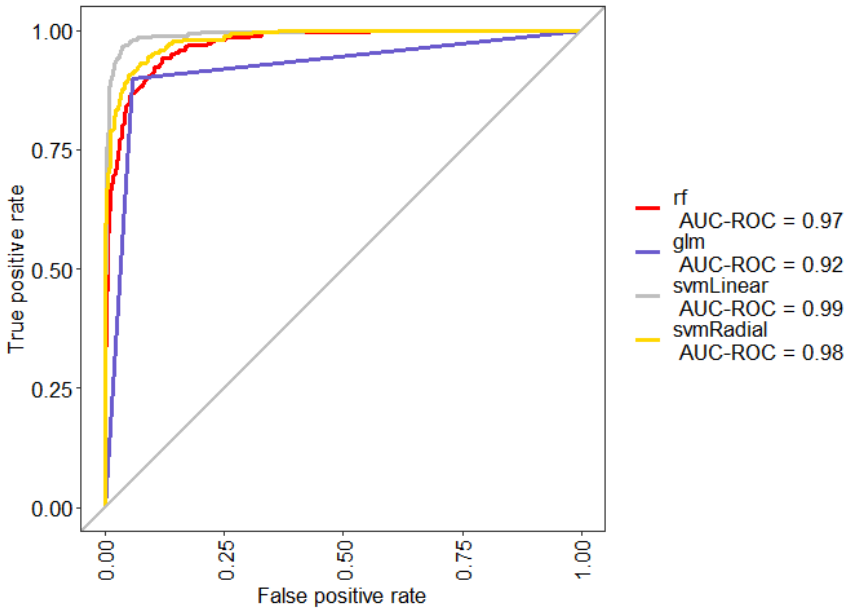


Figure 11. ROC curves for 'complete' models (rf: random forest, glm: logistic regression, svmLinear: SVM with linear kernel, svmRadial: SVM with radial kernel)

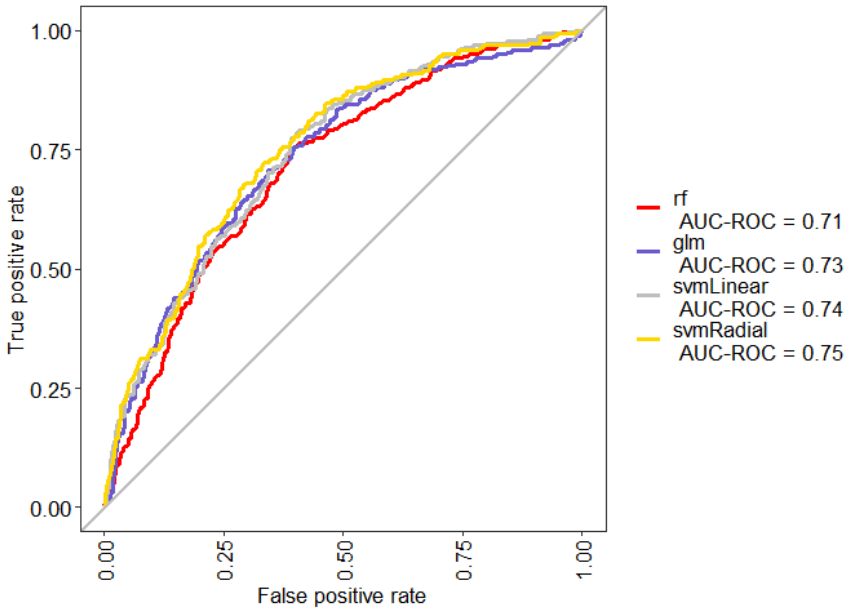


Figure 12. ROC curves for 'reduced' models (rf: random forest, glm: logistic regression, svmLinear: SVM with linear kernel, svmRadial: SVM with radial kernel)

For ease of comparison, performance metrics (sensitivity, specificity, accuracy and kappa statistic) for each model are summarized in Table 10. From this table, the best performing 'complete' model seems to be the SVM model with a radial kernel, although all of them achieved good performance indices even when controlling by the possibility of making correct predictions by chance

(kappa statistic). The kappa statistic was above 0.80 in all cases, which is commonly considered a very good agreement (36). However, performance of the 'reduced' models was generally poor, and although the ROC curves, AUC-ROC values and accuracy indices may indicate an acceptable performance, specificity values were not very different from chance, and the kappa statistics only reached a 'fair' agreement, which may indicate that the models were having difficulties in classifying the cases and probably assigning most cases to the majority class.

Table 10. Performance metrics for the machine learning models for cross-sectional classification.

Model	Sensitivity	Specificity	Accuracy	Kappa
Complete models				
Random forest	0.926	0.913	0.923	0.808
Logistic regression	0.952	0.913	0.942	0.852
SVM – Linear	0.966	0.929	0.956	0.888
SVM – Radial	0.986	0.921	0.969	0.918
Reduced models				
Random forest	0.813	0.524	0.737	0.332
Logistic regression	0.793	0.603	0.743	0.374
SVM – Linear	0.813	0.579	0.752	0.380
SVM – Radial	0.805	0.611	0.754	0.396

4.6 Classification and prediction with longitudinal data

The analysis of longitudinal data was performed by means of (generalized) linear mixed models and with the MEml methods. For this, only complete cases with repeated measures were used, which yielded a total of 3481 observations belonging to 834 different subjects.

4.6.1 Mixed effects generalized linear model: logistic regression

We used a mixed effects logistic regression model to classify cases with or without lipodystrophy while taking into account the repeated measures structure of our dataset. To divide the sample between the training and test sets, we first assigned 75% of the observations to the training set (preserving observations with the same subject ID together). This resulted in 2627 cases, 2008 with lipodystrophy and 619 classified as normal. Given this imbalance, cases with lipodystrophy were randomly removed until achieving class balance in the training set (619 observations in each class, 1238 in total). The test set contained 854 cases, from which 609 were above the cutoff for lipodystrophy and 245 were within the normal range. This way, the classifier is trained in a balanced dataset but is tested with the proportions that are more likely to be observed in a real-world setting.

A logistic regression model was trained with lipodystrophy as categorical outcome (lipodystrophy vs. normal), with age, gender, height, weight and DXA

measures as fixed effects predictors and with the addition of a random effects term corresponding to subject ID. The resulting model showed that gender and trunk fat percent were the only significant predictors of lipodystrophy ($p < 0.05$), with trend-level significance ($p < 0.1$) for left leg fat mass and L1BMD. Performance in the test set showed accuracy = 0.95 (sensitivity = 0.94, specificity = 0.96) and kappa = 0.87. However, the model estimation procedure indicated that the model failed to converge, even though variables had been rescaled and centered, and the proportions of cases in each category were the same in the training set.

One of the possible problems with highly correlated data is multicollinearity (i.e. high correlations between predictors). This can cause the model to fail to converge and make the estimations obtained for the model coefficients unstable, so an alternative model was built in a stepwise fashion. First, univariate models were built for each predictor plus the random effects term. From these, those with p-values < 0.20 were selected, which resulted in a total of 34 variables for the multivariate model. Then, variables were entered in the model by groups, checking convergence of the model for each group: the first group included 'general' variables (age, gender and weight). The second group were the lean mass variables. Here, significant variables were lean mass in left and right arm, both arms together, left and right leg and both legs together. To avoid introducing highly correlated variables in the model, only lean mass for both arms together and lean mass for both legs together were entered.

The third group were BMD variables. Here, 13 variables were selected (L1BMD, L1T, L1Z, L2BMD, L2T, L2Z, NeckFZ, WardsBMD, WardsT, TrochBMD, TrochT, TrochZ, and TotalFZ). The model converged if any of these variables was entered individually, but adding more than one caused convergence to fail. Therefore, only one variable was selected, TotalFZ, which showed the smallest p-value. Finally, ten fat mass variables had been selected from the univariate model (RLFp, RLFg, LLFp, LLFg, BothLFp, BothLFg, TFp, TFg, TotalFp, TotalFg). However, introduction of any of these variables in the model caused it to fail to converge. Therefore, no fat variables were included.

The final model achieved an accuracy = 0.70 (sensitivity = 0.86, specificity = 0.32) and kappa = 0.20 in the test set. This model tended to classify all cases as lipodystrophy, resulting in many false positives.

4.6.2 Mixed effects linear model for prediction of a continuous outcome

The second mixed effects model was a linear regression to predict FMR with a random effects term (subject ID) and the complete set of predictors as fixed effects. For this model, 67% of the cases were assigned to the training set (2361 observations) and the remaining to the test set (1120 observations). FMR ranged from 0.13 to 8.88 (mean = 1.78, SD = 0.86, median = 1.53) in the training set and from 0.22 to 6.17 (mean = 1.72, SD = 0.82, median = 1.49) in the test set. Figure 13 shows the distribution of FMR in both samples.

The resulting model showed that the most significant predictors were fat percent in both legs, fat percent in the trunk and total fat percent (see details in Annex

7). Predictions in the test set showed a MAE = 0.315, MSE = 0.208 and RMSE = 0.456. As shown by Figure 14, predictions were more accurate for low FMR values but became less precise for high values, and predicted values tended to be lower than the observed values.

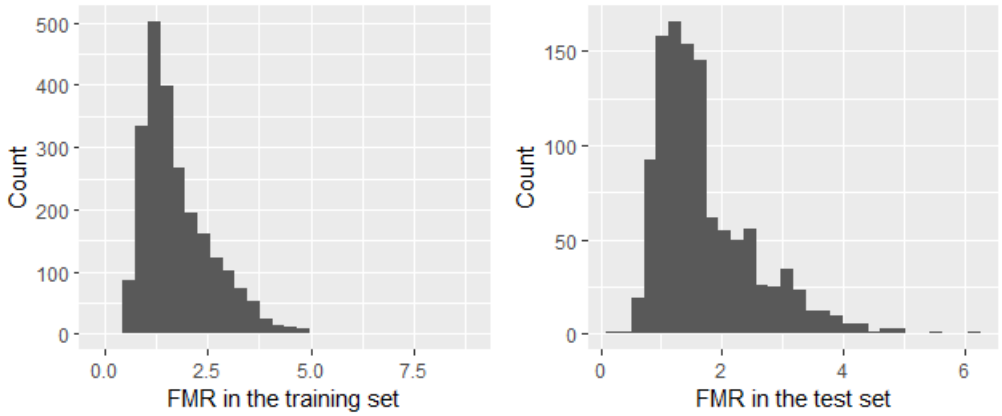


Figure 13. Distribution of FMR in the training and test sets.

4.6.3 Mixed effects machine learning (MEml): prediction of a binary outcome

For the MEml model with a binary outcome, we used the same training/test division as with the mixed effects logistic regression, using the complete set of predictors plus subject ID to identify repeated measures. This model also incorporates a 'visit' variable that chronologically identifies the different observations in each subject. A model was trained with the MEgbm algorithm. On the test set, the model showed sensitivity = 0.75, specificity = 0.80 and AUC-ROC = 0.84.

For comparison, we also trained the 'reduced' model obtained after stepwise introduction of predictors in the mixed effects logistic regression. This one achieved sensitivity = 0.59, specificity = 0.57 and AUC-ROC = 0.62.

4.6.4 Mixed effects machine learning (MEml): prediction of a continuous outcome

The MEml model with a continuous outcome used the same training and test division as the linear regression with mixed effects, and the same predictors and outcome. As with the MEml model with binary outcome, a 'visit' variable was included to sort different observations of the same subject. Here, two models were fitted. The first used the MEgbm algorithm, as above. Performance on the test set showed MAE = 0.519, MSE = 0.403 and RMSE = 0.634. For comparison, we also fit a MErf model, which on the test set showed MAE = 0.440, MSE = 0.340, and RMSE = 0.583. Figure 14 shows that in both models the range of predicted values was very narrow compared to the range of observed values, and predictions tended to overestimate low FMR values (predicted values were higher than observed values) while underestimating high FMR values (predictions were lower than observed values).

4.6.5 Lagged models

The lagged models take advantage of the longitudinal structure of the dataset to predict the outcome variable in the subsequent visit based on the values of the predictors in the current visit. The current value of the outcome was not included as predictor for future visits. The division between training and test was performed automatically by the *LongiLagSplit()* function, which also introduces the lagged structure. This causes the loss of one observation per subject: since the outcome becomes the predicted variable in the next visit, there is no possible prediction for the last visit of each subject. As a result, the dataset ended up with 1992 observations, from which 1627 were assigned to the training set and 365 to the test set.

Two models were fitted, both using the MEgbm algorithm. The first predicted lipodystrophy as a categorical variable (lipodystrophy vs. normal), and showed, on the test set, sensitivity = 0.84, specificity = 0.84 and AUC-ROC = 0.91. The second model predicted FMR as a continuous outcome and showed, on the test set, MAE = 0.294, MSE = 0.219 and RMSE = 0.468. As shown in Figure 14, there was a considerable agreement between predicted and observed values, although predictions seemed to be more accurate (showing less dispersion) for low FMR values than for high FMR values.

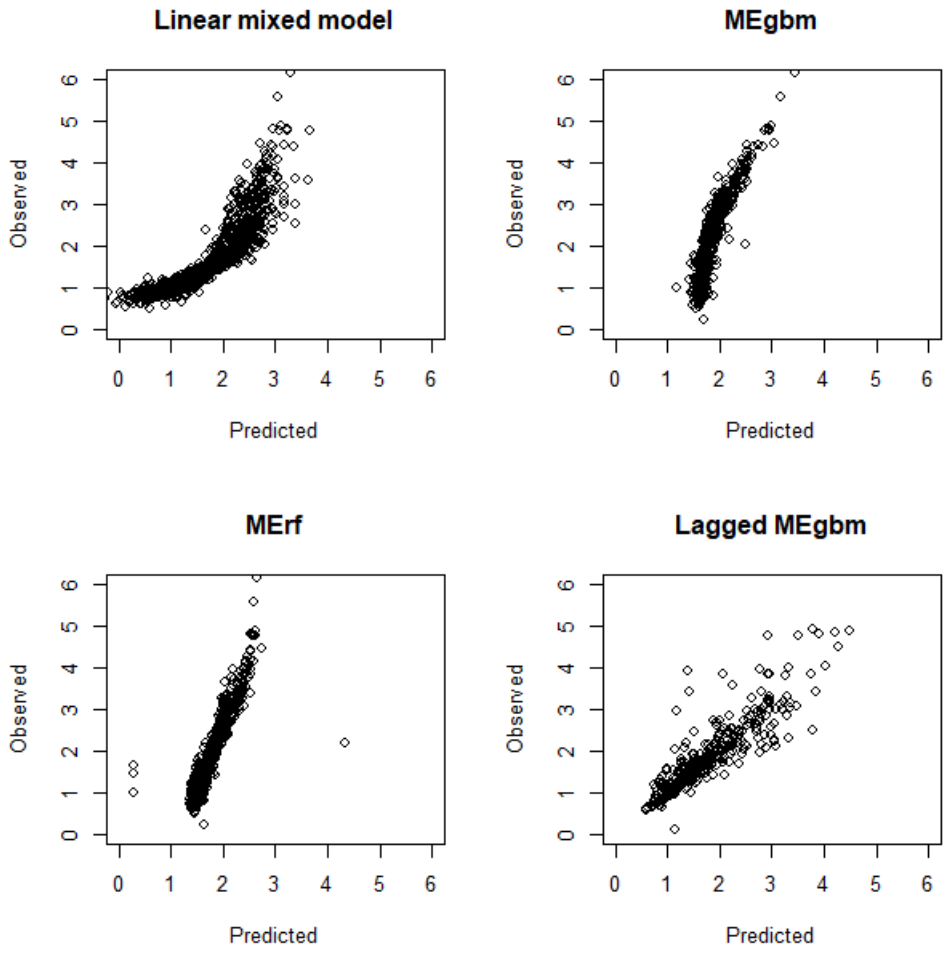


Figure 14. Plots of predicted against observed values for FMR in the longitudinal models with a continuous outcome.

5. Discussion

This work has analyzed a real dataset with the outputs from DXA scans from a large sample of HIV-infected individuals with repeated measurements. The resulting dataset, as is common when working with real data from clinical settings, was unbalanced, with a different number of measurements for each subject, and contained missing values and coding errors, which made necessary a thorough work of data management. For example, as can be seen in Table 1, the number of available cases in each variable was different, and missing values were more common in BMD measures. Since many analysis methods cannot handle missing data, some cases had to be discarded, although sample size was still large and sufficient for the kind of analyses performed here.

As expected, the dataset was highly correlated. The network graphs depicting the correlation structure confirmed that variables coding for the same tissue types clustered together, and this held true even when variable dependencies were studied with a GGM, which was aimed to disentangle direct from indirect effects. However, while Pearson correlations showed that, in general, measures of the same tissue type correlated positively, and fat and lean measures tended to correlate negatively, the GGM revealed some negative associations between measures of the same tissue type, although these were less frequent and generally of lower magnitude than the positive associations. For example, in Figure 4, lean mass of the left and right legs showed a negative association after controlling for the effect of all other variables. This is probably a consequence of both variables correlating with the general variable that coded lean mass for both legs. The graphs also showed the expected negative association between age and BMD measures, and a positive association between age and fat mass measures, that was less evident in the GGM, probably due to the association between age and weight, which was in turn closely associated to lean and fat mass, and would explain an indirect association between age and fat mass or lean mass measures. In general, both correlation approaches converge in a similar correlation structure for the dataset, and highlight the importance of taking it into account when performing classification and prediction.

As noted in the introduction, one common approach for dealing with this kind of data is to use a dimension reduction method like PCA. In the present data, 6 principal components were able to capture more than 90% of the variance in the original 58 variables. A close examination of these components showed that, as would be expected, they captured the variance of groups of variables that were highly correlated in the correlation graphs. For example, BMD measures tended to load into the first component, which could be considered a composite measure of BMD. In the second, fat mass measures loaded positively and lean mass measures loaded negatively, which reflects the negative association between these measures that can be observed in Figure 3. This is interesting not only because it provides insight on the structure and dependencies of the data, but also because these components could be used as summary measures

for describing the trajectories of the patients, similar to some of the indices that are usually calculated (e.g. the body mass index), but potentially richer since they incorporate information from many variables and may capture subtle relationships that the simpler measures cannot incorporate. One example of this may be component 5, which seemed to capture a measure of lipodystrophy, with negative loadings for legs fat mass and fat percentage, and positive loadings for trunk fat, but also incorporated positive loadings (with lower values) for fat mass in the arms and showed a high positive loading for age. On the one hand, this likely reflects the high number of cases with lipodystrophy in this sample, so that most of the observations are characterized by high fat percentage in the trunk and low fat percentage in the legs. On the other hand, it also offers a hint on other variables linked to lipodystrophy that could be incorporated in the diagnosis, and the importance of age as a risk factor for fat mass distribution alterations (2). Thus, scores on this factor could be regarded as a continuous, enriched measure of lipodystrophy, and their trajectories through different measurements across the years could be used to summarize the patients' evolution. However, more research is needed for a clinical validation of such a measure, although its applied potential is worth considering.

The second goal of this project was to use machine learning methods to classify patients according to the presence or absence of lipodystrophy. This was pursued in two ways. In the first, we used only one observation per subject to perform cross-sectional classification based on the measures obtained from the DXA scans. Here two types of model were tested: 'complete' models, that incorporated all the available variables, and 'reduced' models, that aimed to predict lipodystrophy using information from BMD and lean mass, and these models were estimated using different algorithms: random forest, logistic regression and SVM (with linear and with radial kernels). The results showed that all algorithms performed similarly. For the complete models, the best performing model was the SVM with a radial kernel, with a very high accuracy and a kappa statistic above 0.90, although all models achieved accuracies above 0.90 and kappa values above 0.80, with near-perfect AUC-ROC values. This is unsurprising, given that the variables used to diagnose lipodystrophy (legs and trunk fat percent) were included among the predictors. Thus, the performance of the model is probably reflecting how well the algorithm captured the formula for diagnosis. This is confirmed if we examine the variable importance plot in the random forest model, the regression coefficients in logistic regression or the variable weights in the SVM models: all of them show that the most determinant variables are trunk and leg fat mass and fat percent, in some cases accompanied by other fat mass/percent variables. The random forest model also highlights the importance of age.

These complete models achieved better performance than the one proposed in Carr et al. (39), even though they do not include metabolic variables or information regarding HIV infection duration or clinical severity. However, a direct comparison is not possible since this previous model was aimed to predict lipodystrophy based on a clinical diagnosis, not FMR. An interesting goal for future research would be to compare predictions based on FMR and on clinical diagnosis, and also explore the links between DXA-derived measures and metabolic indices.

Model performance dropped when fat mass and fat percent variables were excluded from the predictors. The reduced models show acceptable values for accuracy (above 0.70) but a close examination showed that while sensitivity was good, specificity was poor, and the kappa statistic relatively low. All the reduced models performed similarly, with a tendency to classify cases into the majority class. This resulted in a high rate of false positives, and a drastic reduction in the kappa statistic since most guesses seemed to be due to chance (i.e. classifying subjects in the most frequent category grants more correct guesses, but does not make the model better). Examination of the variable importance plot, coefficients and variable weights showed that, when excluding fat mass variables, age was the most important variable for classification (except for the logistic regression model, which included mainly BMD measures). However, the results of the reduced models indicate that BMD and lean mass measures alone are not enough to accurately classify cases according to the presence or absence of lipodystrophy, despite the links between the three types of measures shown in the correlation analyses and in previous reports (3,34). Therefore, classification and prediction of lipodystrophy should rely, for the moment, in fat mass measures.

The second approach for classification and prediction was to incorporate the longitudinal data into the models. The analysis of repeated measures data has been traditionally performed with mixed linear models, while machine learning methods have rarely incorporated the repeated measures structure into the models. Recent developments are starting to take into account the longitudinal data in data-driven methods, like the MEml framework (42), which was tested here along with classical linear mixed models. First, we aimed to perform classification using a mixed effects generalized linear model with logistic regression, using the complete set of DXA variables as predictors and incorporating a random effects term to account for repeated measures. The resulting model showed good performance indices for prediction, but we obtained a warning that the model failed to converge, which could affect the stability of the estimates for the coefficients, and which seemed to be caused by multicollinearity of the predictors. To address this problem, a new model was built in which we tried not to include variables that were too correlated between them. Here, the inclusion of fat mass measures caused the model to fail to converge, so we obtained a reduced model that included age, gender, weight, lean mass of both arms, lean mass of both legs and the Z-score of total femur BMD. However, this model, like the reduced cross-sectional models, showed poor predictive capacity, with a very high rate of false positives. The same models were also estimated with the MEgbm algorithm, which showed good performance with the complete set of predictors, although slightly lower than the mixed effects logistic regression. However, estimation with the machine learning approach did not result in problems of model convergence, so estimates might be ultimately more reliable in this latter case. With the predictors selected in the stepwise model construction, the specificity and sensitivity values obtained were, again, rather poor. That is, classification seemed to be more dependent on the selected predictors than the specific algorithm, as happened with cross-sectional classification.

Besides classification, we also used the longitudinal data for prediction of a continuous outcome, FMR. This is of interest, on the one hand, because instead of relying in an absolute threshold, FMR also indicates how far from the threshold the patient is, i.e. the severity of lipodystrophy, and on the other hand because FMR has been associated with other metabolic outcomes linked to cardiovascular risk. In parallel with the models with binary outcomes, we estimated a mixed effects linear regression model, and two MEml models (using the MEgbm and the MErf algorithms). In this case, models were compared based on the prediction errors, and the mixed effects linear regression showed the lowest values for mean square error, mean absolute error and root mean square error. Therefore, it seems that for prediction of FMR based on DXA measures, parametric models are more powerful. Accordingly, the plots in Figure 14 also showed that predicted values in the linear mixed effects model were more accurate than those observed with the MEgbm and MErf algorithms.

Finally, one of the most interesting insights to extract from longitudinal data is the prediction of future values of the outcome variable based on present and past measurements. We tested whether we could apply the MEml tools in this dataset to predict the presence/absence of lipodystrophy and the values of FMR in the next visit, based on the DXA measures of the present one. The model predicting presence of lipodystrophy showed very good sensitivity and specificity values, indicating that it was possible to predict lipodystrophy one visit ahead. Importantly, the information on whether or not the patient showed lipodystrophy in the present visit was not included as a predictor, so predictions were based only on DXA values. Similarly, Figure 14 also shows that prediction of FMR in the next visit was considerably accurate, with a considerable and approximately linear agreement between predicted and observed values, although predictions for cases with high FMR tended to be less accurate.

Therefore, this work has demonstrated the feasibility of such predictions, which have potential application in the monitoring and prognosis of HIV-infected individuals. Here, due to the available longitudinal data, prediction was limited to one visit ahead. However, with the accumulation of more information and more complete datasets, the predictions can be extended to assessments further in the future. The framework is still in development, so future improvements can be expected. For example, a more accurate definition of ‘visit’ that takes into account the actual time elapsed between scans could aid in making more accurate predictions.

The present work shows the potential of unsupervised and supervised data-driven methods to derive useful insights and potential applications from clinical, real-world data. However, it also manifests the importance of considering the repeated measures structure when analyzing this type of data, a feature that the most common machine learning methods have not usually incorporated. In fact, our results show that in some cases, parametric approaches may show more powerful results (e.g. if assumptions are met). We have also shown the potential of predicting how a patient is going to evolve based on their past trajectory. This may be useful in the identification of cases at high risk of developing cardiovascular conditions or metabolic syndrome. The use of DXA measures alone to make the predictions may pose a limitation: lipodystrophy is

also characterized by metabolic alterations, but no metabolic indices were available in this dataset. On the other hand, it has been noted that the increase in trunk fat in lipodystrophy refers to visceral fat, while subcutaneous fat in the abdomen could be actually reduced (11). DXA scans cannot differentiate visceral fat from subcutaneous fat. This has to be estimated by computerized tomography (CT) or, preferentially, by magnetic resonance imaging (MRI), but this is only commonly used in research settings. Our approach uses data from clinical settings, so it is limited to DXA, but at the same time is more generalizable, because DXA is performed more frequently than CT or MRI. Moreover, the most recent DXA scanners are starting to incorporate ways of measuring visceral fat and subcutaneous fat separately (21). Therefore, the predictive methods that have been explored in this project should be able to incorporate these recent developments to obtain accurate predictive tools aimed at better diagnosis, prognosis and prevention of health complications associated to HIV infection.

6. Conclusions

This work has explored the application of machine learning methods for classification and prediction of lipodystrophy in HIV-infected individuals. In general, we can extract the following conclusions:

- The large number of outputs from the DXA-scans can be summarized in a few principal components that relate to different aspects of body composition, with components that seem specifically linked to osteoporosis, low muscle mass and lipodystrophy.
- Machine learning methods can accurately discriminate patients with and without lipodystrophy in a cross-sectional analysis, but only if fat mass measures are included among the predictors. Classification is poor with only lean mass and BMD measures.
- Parametric mixed effects models can accurately predict lipodystrophy or FMR, although multicollinearity may interfere in model estimation.
- The combined MEml approach was successful in classifying by presence or absence of lipodystrophy using the repeated measures structure, although prediction of FMR was more accurate with the parametric approach.
- MEml can be used for longitudinal prediction of future lipodystrophy status or FMR values.

The greatest challenge in this project has been working with real data with high complexity that required, on the one hand, a thorough and accurate process of data management to identify missing values, coding errors, and duplicates, plus some problems caused by how Excel treats empty cells that required recalculation of variables that were composite measures from other variables. On the other hand, dealing with a highly correlated dataset has also required a thorough exploration of the dependencies between the variables and the consideration of the repeated measures structure.

In general, the project goals have been achieved, although some of the results have been negative. For example, based on prior evidence it was expected that some kind of predictive value for lipodystrophy could be obtained from lean mass and BMD measures, but the results have shown that fat mass measures were necessary for this prediction to be accurate. On the other hand, the outcomes of the work are still some steps away from a real-world application. However, some of the results are potentially interesting and could lead to future developments of new summary measures (based e.g. on PCA) or predictive tools.

In terms of adherence to the work plan, it has been followed during most part of the project's development, although some slight changes had to be introduced to adapt the work to the intermediate findings and the available tools. The most important change has been the introduction of linear mixed models, which was not contemplated in the initial planning but proved to be a useful and interesting way to analyze repeated measures data and provided the grounds for

comparison with the more novel MEml methods. Given the novelty of the tools, and the scarcity of machine learning methods for repeated measures, the longitudinal analysis took longer than expected. However, in general the initial work plan has been followed with very small changes and all planned analyses have been run.

Finally, this work has some limitations that may be overcome in future developments. An interesting follow-up of the present results would be to explore the potential of PCA scores to act as summary measures. Considering that these are data-driven measures obtained from a large number of variables (instead of the currently used summary measures that are ratios or combinations of two or three values), they could represent more complete, richer measures. For this, we should examine how these components behave, what characterizes patients with high or low scores, and their associations with other clinical outcomes other than DXA measures. Other possible descriptive tools can also be considered. For example, the analysis of proximities or the application of cluster analysis may help to identify patient subgroups, and the examination of these subgroups may identify specific clinical profiles. Another, very promising future research line would be to develop tools to accurately predict the evolution of FMR/lipodystrophy, which could be also extended to BMD or lean mass measures. This work has shown that it is possible to predict, quite accurately, the presence of lipodystrophy or the values of FMR in a future measurement, based on current and past data. Therefore, future developments should explore to which degree these predictions can be extended to anticipate the patients' trajectories and prevent future health complications.

7. Glossary

HIV: Human Immunodeficiency Virus. HIV is a virus from the *Lentivirus* genus, part of the *Retroviridae* family, that infects humans. HIV infects immune cells ($CD4^+$ T cells, macrophages, and dendritic cells), leading to low and impaired immune function and, if untreated, immunodeficiency.

Lipodystrophy: An alteration of body fat distribution that mainly consists of a reduction of subcutaneous fat in the face and extremities, and/or an increase in visceral fat in the trunk. It can also be accompanied by metabolic alterations.

cART: Combined Antiretroviral Therapy. Pharmacological treatment for HIV infection that uses three or more drugs. Also called Highly Active Antiretroviral Therapy (**HAART**).

DXA: Dual-energy X-ray absorptiometry. Medical imaging technique that uses spectral imaging to measure Bone Mineral Density and body composition (differentiating lean mass from fat mass).

Machine learning: Scientific discipline in the field of artificial intelligence aimed at the creation of systems that learn automatically.

GGM: Gaussian Graphical Model. A type of statistical model that explicitly captures the statistical relationships between the variables of interest and represents them in graph form. Dependencies between variables are computed as full-order partial correlations.

PCA: Principal Component Analysis. Data reduction technique that finds a set of principal components, which are linear combinations of the original variables, that explain most of the original variance in the first few principal components. Principal components are uncorrelated and ordered according to the proportion of the original variance that they explain.

Random forest: Classification model that combines multiple decision trees working as an ensemble, with the output class being the most frequently chosen class in the ensemble.

Logistic regression: Type of regression model that is used to predict the result of a categorical variable. The output is the probability of belonging to one of the classes.

SVM: Support Vector Machine. Supervised learning model that maps training examples to points in an n-dimensional space (n = number of predictors) so as to maximize separation between categories. New examples are mapped into that same space and predicted to belong to a category depending on where they fall relative to the separation boundary.

Mixed models: Statistical models that contain both fixed and random effects. They are particularly useful to model data with repeated measures.

GBM: Gradient boosting machine. Supervised learning technique for regression and classification that creates a prediction model as an ensemble of weak prediction models, typically decision trees.

MEml: Mixed Effects machine learning. Approach to the study of longitudinal data through machine learning algorithms that incorporate the repeated measures structure.

MEgbm: Mixed Effects gradient boosting machine. Machine learning algorithm based on the GBM that incorporates repeated measures.

MErf: Mixed Effects random forest. Machine learning algorithm based on random forest that incorporates repeated measures.

8. References

1. The Antiretroviral Therapy Cohort Collaboration. Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies. Lancet [Internet]. 2008 Jul;372(9635):293–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673608611137
2. Koethe JR, Lagathu C, Lake JE, Domingo P, Calmy A, Falutz J, et al. HIV and antiretroviral therapy-related fat alterations. Nat Rev Dis Prim [Internet]. 2020;6(1). Available from: http://dx.doi.org/10.1038/s41572-020-0181-1
3. Buehring B, Kirchner E, Sun Z, Calabrese L. The Frequency of Low Muscle Mass and Its Overlap With Low Bone Mineral Density and Lipodystrophy in Individuals With HIV-A Pilot Study Using DXA Total Body Composition Analysis. J Clin Densitom [Internet]. 2012;15(2):224–32. Available from: http://dx.doi.org/10.1016/j.jocd.2011.10.003
4. McDermott AY, Shevitz A, Knox T, Roubenoff R, Kehayias J, Gorbach S. Effect of highly active antiretroviral therapy on fat, lean, and bone mass in HIV-seropositive men and women. Am J Clin Nutr. 2001;74(5):679–86.
5. Borderi M, Gibellini D, Vescini F, De Crignis E, Cimatti L, Biagetti C, et al. Metabolic bone disease in HIV infection. Aids. 2009;23(11):1297–310.
6. Kanis JA, Johnell O, Oden A, Dawson A, De Laet C, Jonsson B. Ten year probabilities of osteoporotic fractures according to BMD and diagnostic thresholds. Osteoporos Int. 2001;12(12):989–95.
7. Kanis JA, McCloskey E V., Johansson H, Oden A, Melton LJ, Khaltaev N. A reference standard for the description of osteoporosis. Bone. 2008;42(3):467–75.
8. Rosenberg IH. Sarcopenia: Origins and clinical relevance. Clin Geriatr Med. 2011;27(3):337–9.
9. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, et al. Sarcopenia: Revised European consensus on definition and diagnosis. Age Ageing. 2019;48(1):16–31.
10. Gould H, Brennan SL, Kotowicz MA, Nicholson GC, Pasco JA. Total and appendicular lean mass reference ranges for Australian men and women: The Geelong osteoporosis study. Calcif Tissue Int. 2014;94(4):363–72.
11. Freitas P, Santos AC, Carvalho D, Pereira J, Marques R, Martinez E, et al. Fat Mass Ratio: An Objective Tool to Define Lipodystrophy in HIV-Infected Patients Under Antiretroviral Therapy. J Clin Densitom. 2010;13(2):197–203.
12. Podzamczer D, Ferrer E, Martínez E, Del Rio L, Rosales J, Curto J, et al. How much fat loss is needed for lipoatrophy to become clinically evident? AIDS Res Hum Retroviruses. 2009;25(6):563–7.
13. Messina C, Monaco CG, Ulivieri FM, Sardanelli F, Sconfienza LM. Dual-energy X-ray absorptiometry body composition in patients with secondary osteoporosis. Eur J Radiol [Internet]. 2016;85(8):1493–8. Available from: http://dx.doi.org/10.1016/j.ejrad.2016.03.018
14. Everitt B, Hothorn T. An Introduction to Applied Multivariate Analysis with R [Internet]. An Introduction to Applied Multivariate Analysis with R. New York, NY: Springer New York; 2011. Available from:

http://link.springer.com/10.1007/978-1-4419-9650-3
15. Altenbuchinger M, Weihs A, Quackenbush J, Grabe HJ, Zacharias HU. Gaussian and Mixed Graphical Models as (multi-)omics data analysis tools. Biochim Biophys Acta - Gene Regul Mech [Internet]. 2020;1863(6):194418. Available from: https://doi.org/10.1016/j.bbagrm.2019.194418
16. Sedgewick AJ, Buschur K, Shi I, Ramsey JD, Raghu VK, Manatakis D V., et al. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. Bioinformatics. 2019;35(7):1204–12.
17. World Health Organization. HIV/AIDS [Internet]. 2020. Available from: https://www.who.int/news-room/fact-sheets/detail/hiv-aids
18. Brown TT, Xu X, John M, Singh J, Kingsley LA, Palella FJ, et al. Fat distribution and longitudinal anthropometric changes in HIV-infected men with and without clinical evidence of lipodystrophy and HIV-uninfected controls: A substudy of the Multicenter AIDS Cohort Study. AIDS Res Ther. 2009;6:1–8.
19. Guaraldi G, Stentarelli C, Zona S, Santoro A. HIV-associated lipodystrophy: Impact of antiretroviral therapy. Drugs. 2013;73(13):1431–50.
20. Bonnet E, Delpierre C, Sommet A, Marion-Latard F, Hervé R, Aquilina C, et al. Total Body Composition by DXA of 241 HIV-Negative Men and 162 HIV-Infected Men. J Clin Densitom [Internet]. 2005 Sep;8(3):287–92. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1094695006603193
21. Messina C, Albano D, Gitto S, Tofanelli L, Bazzocchi A, Ulivieri FM, et al. Body composition with dual energy X-ray absorptiometry: From basics to new tools. Quant Imaging Med Surg. 2020;10(8):1687–98.
22. Vasandani C, Li X, Sekizkardes H, Adams-Huet B, Brown RJ, Garg A. Diagnostic Value of Anthropometric Measurements for Familial Partial Lipodystrophy, Dunnigan Variety. J Clin Endocrinol Metab [Internet]. 2020 Jul 1;105(7):2132–41. Available from: https://academic.oup.com/jcem/article/105/7/2132/5810271
23. Dubé MP, Komarow L, Mulligan K, Grinspoon SK, Parker RA, Robbins GK, et al. Long-term body fat outcomes in antiretroviral-naive participants randomized to nelfinavir or efavirenz or both plus dual nucleosides: Dual x-ray absorptiometry results from A5005s, a substudy of adult clinical trials group 384. J Acquir Immune Defic Syndr. 2007;45(5):508–14.
24. Dubé MP, Parker RA, Tebas P, Grinspoon SK, Zackin RA, Robbins GK, et al. Glucose metabolism, lipid, and body fat changes in antiretroviral-naive subjects randomized to nelfinavir or efavirenz plus dual nucleosides. Aids. 2005;19(16):1807–18.
25. Gundurao Sreekantamurthy G, Singh NB, Singh TB, Singh TS, Singh KR. Study of body composition and metabolic parameters in hiv-1 male patients. J Nutr Metab. 2014;2014(September 2013):1–6.
26. Lichtenstein K a. Redefining Lipodystrophy Syndrome Risks and Impact on Clinical Decision Making. J Acquir Immune Defic Syndr. 2005;39(4):395–400.
27. Grunfeld C, Rimland D, Gibert CL, Powderly WG, Sidney S, Shlipak MG, et al. Association of Upper Trunk and Visceral Adipose Tissue Volume

With Insulin Resistance in Control and HIV-Infected Subjects in the FRAM Study. JAIDS J Acquir Immune Defic Syndr [Internet]. 2007 Nov;46(3):283–90. Available from: http://journals.lww.com/00126334-200711010-00005
28. Wohl D, Scherzer R, Heymsfield S, Simberkoff M, Sidney S, Bacchetti P, et al. The Associations of Regional Adipose Tissue With Lipid and Lipoprotein Levels in HIV-Infected Men. JAIDS J Acquir Immune Defic Syndr [Internet]. 2008 May;48(1):44–52. Available from: http://journals.lww.com/00126334-200805010-00006
29. Nieves DJ, Cnop M, Retzlaff B, Walden CE, Brunzell JD, Knopp RH, et al. The Atherogenic Lipoprotein Profile Associated With Obesity and Insulin Resistance Is Largely Attributable to Intra-Abdominal Fat. Diabetes [Internet]. 2003 Jan 1;52(1):172–9. Available from: http://diabetes.diabetesjournals.org/cgi/doi/10.2337/diabetes.52.1.172
30. Carr DB, Utzschneider KM, Hull RL, Kodama K, Retzlaff BM, Brunzell JD, et al. Intra-Abdominal Fat Is a Major Determinant of the National Cholesterol Education Program Adult Treatment Panel III Criteria for the Metabolic Syndrome. Diabetes [Internet]. 2004 Aug 1;53(8):2087–94. Available from: http://diabetes.diabetesjournals.org/cgi/doi/10.2337/diabetes.53.8.2087
31. Calvo M, Martinez E. Update on metabolic issues in HIV patients. Curr Opin HIV AIDS. 2014;9(4):332–9.
32. Freitas P, Carvalho D, Santos AC, Mesquita J, Matos MJ, Madureira AJ, et al. Lipodystrophy defined by Fat Mass Ratio in HIV-infected patients is associated with a high prevalence of glucose disturbances and insulin resistance. BMC Infect Dis [Internet]. 2012 Dec 6;12(1):180. Available from: https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-12-180
33. Kristoffersen US, Lebech A-M, Wiinberg N, Petersen CL, Hasbak P, Gutte H, et al. Silent Ischemic Heart Disease and Pericardial Fat Volume in HIV-Infected Patients: A Case-Control Myocardial Perfusion Scintigraphy Study. Emery S, editor. PLoS One [Internet]. 2013 Aug 14;8(8):e72066. Available from: https://dx.plos.org/10.1371/journal.pone.0072066
34. Yao J, Yu W, Li T, Luo L, Lin Q, Tian J, et al. The Pilot Study of DXA Assessment in Chinese HIV-Infected Men With Clinical Lipodystrophy. J Clin Densitom [Internet]. 2011;14(1):58–62. Available from: http://dx.doi.org/10.1016/j.jocd.2010.08.002
35. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph : Network Visualizations of Relationships in Psychometric Data. J Stat Softw [Internet]. 2012;48(4). Available from: http://www.jstatsoft.org/v48/i04/
36. Lantz B. Machine Learning with R: Second Edition. Machine Learning with R. 2015.
37. Yiu T. Understanding Random Forest. How the Algorithm Works and Why it Is So Effective [Internet]. 2019 [cited 2020 Dec 22]. Available from: https://towardsdatascience.com/understanding-random-forest-58381e0602d2
38. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning [Internet]. New York, NY: Springer New York; 2013. (Springer Texts in Statistics; vol. 103). Available from:

http://link.springer.com/10.1007/978-1-4614-7138-7
39. Carr A, Emery S, Law M, Puls R, Lundgren JD, Powderly WG, et al. An objective case definition of lipodystrophy in HIV-infected adults: A case-control study. Lancet. 2003;361(9359):726–35.
40. UCLA: Statistical Consulting Group. Introduction to generalized linear mixed models [Internet]. [cited 2020 Dec 4]. Available from: https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/
41. Heagerty PJ. Analysis of Correlated Data [Internet]. 2006 [cited 2020 Nov 4]. Available from: https://faculty.washington.edu/heagerty/Courses/VA-longitudinal/private/CorrData-Intro.pdf
42. Ngufor C, Van Houten H, Caffo BS, Shah ND, McCoy RG. Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. J Biomed Inform [Internet]. 2019;89:56–67. Available from: https://doi.org/10.1016/j.jbi.2018.09.001
43. Grolemund G, Wickham H. Dates and Times Made Easy with lubridate. J Stat Softw [Internet]. 2011;40(3). Available from: http://www.jstatsoft.org/v40/i03/
44. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw [Internet]. 2008;28(5). Available from: http://www.jstatsoft.org/v28/i05/
45. Kuhn M. The caret Package [Internet]. 2019 [cited 2020 Oct 5]. Available from: https://topepo.github.io/caret/
46. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res [Internet]. 2002 Jun 1;16:321–57. Available from: https://www.jair.org/index.php/jair/article/view/10302
47. Bates D, Maechler M, Bolker B, Walker S, Haubo R, Christensen B, et al. Package lme4: Linear Mixed-Effects Models using “Eigen” and S4 [Internet]. 2020. Available from: https://cran.r-project.org/web/packages/lme4/lme4.pdf

9. Annexes

Annex 1. Supplementary tables.
File: Annex 1 Supplementary tables.docx

Annex 2. Original dataset
File: Totes DEXES completes_09-01-18.xlsx

Annex 3. Data preparation code and outputs
File: db_prep_rev.zip (contains db_prep_rev.Rmd and db_prep_rev.html)

Annex 4. Curated dataset
File: dxa.csv

Annex 5. Code and outputs for correlation analyses and PCA
File: corr_pca.zip (contains corr_pca.Rmd and corr_pca.html)

Annex 6. Code and outputs for cross-sectional machine learning models
File: cross_sectional.zip (contains cross_sectional.Rmd and cross_sectional.html)

Annex 7. Code and outputs for longitudinal machine learning models
File: longitudinal.zip (contains longitudinal.Rmd and longitudinal.html)