INTERACTIVE ANALYTICS AND VISUALIZATION FOR DATA DRIVEN CALCULATION OF INDIVIDUALIZED COPD RISK

Master Thesis performed at AMRA AB and Linköping University in collaboration with Wolfram Mathcore AB

January 2018 - June 2018

Master Student/Author: Emil Arkstål

Program: Engineering Biology, M Sc in Engineering

Examiner: Magnus Bång (Department of Computer and Information science (IDA) and Human Centered Systems (HCS))

Supervisors: Mikael Forsgren (AMRA AB and Wolfram Mathcore AB), Olof Dahlqvist Leinhard (AMRA AB), Jennifer Linge (AMRA AB)
# CONTENTS

1 Introduction                                      1
  1.1 Project Rationale                             1
  1.2 Aim                                           1
  1.3 Demarcations                                  1
  1.4 Parameters and abbreviations                  2

2 Background                                       4
  2.1 Chronic obstructive pulmonary disease         4
  2.2 BMI and anthropometric surrogate parameters   6
  2.3 Body composition profiling                    6
  2.4 Software                                      9
  2.5 Graphical Considerations                      10

3 Method                                           13
  3.1 Method overview                               13
  3.2 Implementation of R into WL                  14
  3.3 The data                                      14
  3.4 Creating the virtual control group            16
  3.5 Developing Visualization                     17
  3.6 VCG-based COPD diagnostics                    18
  3.7 Exploring COPD phenotypes                     19

4 Result                                           20
  4.1 Developing visualizations                     20
  4.2 The IICE algorithm                            24
  4.3 VCG-based COPD diagnostics                    28
  4.4 The IICE tool                                 32
  4.5 Characterizing COPD                           36
  4.6 Individual example analysis                   53

5 Discussion                                       58
  5.1 Results                                       58
  5.2 Method                                        64
5.3 Graphical consideration 65
5.4 Future work 67

6 Conclusion 68

References 69

APPENDIX A: Scatter plots of Type A-D i
APPENDIX B: BCPs from data sets x
ABSTRACT

Chronic obstructive pulmonary disease (COPD) is a high mortality disease, second to stroke and ischemic heart disease. This non-curable disease progressively exacerbates, leading to high personal and societal economic impact, reduced quality of life and often death. General treatment plans for COPD risk mistreating the individuals’ condition. To be effective, the treatment should be individualized following the practices of precision medicine.

The aim of this thesis was to develop a data driven algorithm and system with visualization to assess individual COPD risk. With MRI body composition profile measurements, it is possible to accurately assess propensity of a multitude of metabolic conditions, such as coronary heart disease and type 2 diabetes.

The algorithm and system has been developed using Wolfram Language and R within the Wolfram Mathematica framework. The algorithm calculates individualized virtual control groups metabolically similar to the patient’s body composition and spirometric profile. Using UK Biobank data, our tool was used to assess patient COPD propensity using an individual-specific virtual control group with AUROC 0.778 (female) and 0.758 (men). Additionally, the tool was used to identify new body composition profiles related to COPD and associated comorbid conditions.
1 INTRODUCTION

1.1 PROJECT RATIONALE
The World Health Organization estimates that there are currently 65 million people suffering from chronic obstructive pulmonary disease (COPD). Being responsible for 5% of global death in 2005, COPD is a non-curable comorbid disease with a progressively increasing impact on quality of life and both individual and societal cost. Its prevalence is projected to increase by 30% in the next ten years (1).

COPD is a breathing impairing disease, comprising two main conditions, emphysema and chronic bronchitis. It is often accompanied by various comorbidities such as sarcopenia, anorexia, cachexia and osteoporosis. The disease’s faceted, metabolically connected nature leads to different outlooks based on individual body composition (2).

A common risk assessment metric of metabolic disease is BMI. On a population scale higher BMI does correlate well with increased disease risk, but metabolism is complex on the individual level. Two individuals of the same BMI might be almost metabolically incomparable when it comes to body weight distribution, both through muscle to fat ratio but also when comparing their respective fat compartment ratios. The distribution of fat is just as, if not more, important as the total amount of fat when it comes to assessing individual metabolic disease risk (3).

Through MRI scans, individual compartments of fat can be measured, providing more valuable information of ectopic fat rather than overall fat. Using such data, together with spirometric and survey data, precise individual risk assessment can be done. Such a data set was available for this project, containing 6021 subjects.

To utilize the data available through extensive MRI, spirometry and surveys, data visualization and interpretation is essential. The data must be easily accessed and interacted with, allowing the user to quickly extract the desired knowledge, without risk for skewed results. The user should be able to analyze groups who are similar in one aspect by viewing their spread in others. All visualizations should focus on the individual, however, and put them in the context of metabolically similar groups of people.

1.2 AIM
The aim of this project is to develop an algorithm and system with which to assess individual COPD risk and characteristics. The tool will provide insight to individual COPD propensity based on values acquired through magnetic resonance imaging and spirometry. It is also desired for the tool to work well when exploring and assessing other conditions.

There is one main question asked in this project; is it possible to develop an accurate, informative visualization of COPD risk using virtual control group-technology? To answer the question, three sub-questions need analysis:

➢ Can COPD presence be assessed through the available data?
➢ Do any available comorbidities indicate different phenotypes in COPD patients?
➢ What is a suitable visualization?

1.3 DEMARCATIONS
The project comprises a 30 credits course and so has been limited to 20 working weeks (40 hours each). Because of this, within the confines of the project the visualization and analytics tool will be used only to analyze the disease COPD, although it is desired to perform well when analyzing other conditions.
In this report there are several images displaying various graphics copied from the tool. In some cases the labels, legends or other explanatory text might be smaller due to downscaling. Several of the graphics, and their text, are displayed in greater scale within the tool. In these cases the images have the purpose to conceptualize the graphic, not to express interpretable results.

The data available limits the precision of VCG creation. The UK Biobank data contains 109 COPD cases, having been reported by the patients themselves. The COPD vitamin D deficiency study patients do not have complete BCP variable data sets, and as such may not be usable to validate certain aspects of the diagnostic performance.

1.4 Parameters and Abbreviations

Below is a table of relevant variables and their corresponding abbreviations and other abbreviations used in this project.

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Description</th>
<th>Formula or content</th>
</tr>
</thead>
<tbody>
<tr>
<td>aid</td>
<td>AMRA ID</td>
<td></td>
</tr>
<tr>
<td>UKBB</td>
<td>UK Biobank</td>
<td></td>
</tr>
<tr>
<td>VAT</td>
<td>Visceral Adipose Tissue</td>
<td>MRI-measured</td>
</tr>
<tr>
<td>VATi</td>
<td>VAT index</td>
<td>VAT/ height²</td>
</tr>
<tr>
<td>ASAT</td>
<td>Abdominal Subcutaneous Adipose Tissue</td>
<td>MRI-measured</td>
</tr>
<tr>
<td>ASATi</td>
<td>ASAT index</td>
<td>ASAT/height²</td>
</tr>
<tr>
<td>IMAT</td>
<td>IntraMuscular Adipose Tissue or Muscle Fat Infiltration</td>
<td>MRI-measured</td>
</tr>
<tr>
<td>ATAT</td>
<td>Abdominal total adipose tissue</td>
<td>VAT+ASAT</td>
</tr>
<tr>
<td>ATATi</td>
<td>ATAT index</td>
<td>ATAT/height²</td>
</tr>
<tr>
<td>rulf</td>
<td>Right upper leg front muscle</td>
<td>MRI-measured</td>
</tr>
<tr>
<td>lulf</td>
<td>Left upper leg front muscle</td>
<td>MRI-measured</td>
</tr>
<tr>
<td>rulb</td>
<td>Right upper leg back muscle</td>
<td>MRI-measured</td>
</tr>
<tr>
<td>lulb</td>
<td>Left upper leg back muscle</td>
<td>MRI-measured</td>
</tr>
<tr>
<td>MR</td>
<td>Muscle ratio or Weight-to-Muscle Ratio</td>
<td>Weight/(lulb + rulb + lulf + rulf)</td>
</tr>
<tr>
<td>WFR</td>
<td>Weight-to-fat ratio</td>
<td>Weight/(ATAT + VAT)</td>
</tr>
<tr>
<td>FR</td>
<td>Fat ratio</td>
<td>ATAT/(ATAT-(lulb+rulb+rulf+lulf))</td>
</tr>
<tr>
<td>Lff10p</td>
<td>Liver proton density fat fraction (mean of 9 liver regions of interest). In short, liver fat.</td>
<td>Measured through MRI</td>
</tr>
<tr>
<td>BCP</td>
<td>Body Composition Profile</td>
<td>An individuals’ set of values for ATATi, FR, MR, IMAT, VATi and Lff10p</td>
</tr>
<tr>
<td>HCB</td>
<td>Health Care Burden</td>
<td>Number of recorded hospital nights</td>
</tr>
<tr>
<td>Hcb_trunc15</td>
<td>HCB truncated to a maximum of 30 for the last 15 years prior to scan, corrected for pregnancy-related visits.</td>
<td></td>
</tr>
<tr>
<td>Prop_hcb_bcp</td>
<td>Propensity of HCB</td>
<td>See 2.3.2</td>
</tr>
<tr>
<td>TTVi</td>
<td>Total Thigh muscle Volume Index</td>
<td>Calculated from leg muscle variables and height.</td>
</tr>
<tr>
<td>FEV1</td>
<td>Forced Expiratory Volume in the first second</td>
<td>Spirometry-measured</td>
</tr>
<tr>
<td>FVC</td>
<td>Forced Vital Capacity</td>
<td>Spirometry-measured</td>
</tr>
<tr>
<td>lungfun</td>
<td>Lung function</td>
<td>FEV₁/FVC</td>
</tr>
<tr>
<td>---------</td>
<td>---------------</td>
<td>----------</td>
</tr>
<tr>
<td>srCOPD</td>
<td>Self-reported COPD</td>
<td>1 or 0</td>
</tr>
<tr>
<td>Soft copd</td>
<td>Subjects with lungfun&lt;0.7</td>
<td>1 or 0</td>
</tr>
<tr>
<td>T2D</td>
<td>Type 2 Diabetes</td>
<td>1 or 0</td>
</tr>
<tr>
<td>CHD</td>
<td>Coronary Heart Disease</td>
<td>1 or 0</td>
</tr>
<tr>
<td>ppsrCOPD</td>
<td>Propensity for self-reported COPD</td>
<td>Share of srCOPD cases in virtual control group</td>
</tr>
<tr>
<td>ppsarcopenia</td>
<td>Propensity for MRI- and DXA-diagnosed sarcopenia, which is based on a cut-off value for TTVi</td>
<td>Share of sarcopenic cases in virtual control group</td>
</tr>
<tr>
<td>ppT2D</td>
<td>Propensity for Type-2-diabetes</td>
<td>Previously calculated share of T2D cases in a virtual control group created using BCP variables</td>
</tr>
<tr>
<td>ppCHD</td>
<td>Propensity for Coronary Heart Disease</td>
<td>Previously calculated share of CHD cases in a virtual control group created using BCP variables</td>
</tr>
<tr>
<td>ppHCB</td>
<td>Propensity for HCB</td>
<td>Previously calculated level of HCB in a virtual control group created using BCP variables</td>
</tr>
</tbody>
</table>
2 BACKGROUND

2.1 CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Chronic obstructive pulmonary disease (COPD) is a disease that progressively impairs breathing capabilities. In 2015 it was one of the main causes of death globally, and currently there is no cure (4). The progression rate of COPD varies from patient to patient but is most often irreversible. As respiratory function deteriorates, COPD causes dyspnea for progressively menial activities, eventually even eating (2).

COPD generally comprises two main conditions, emphysema and chronic bronchitis, which can occur both together and individually. Emphysema is a condition in which the alveoli walls are damaged, causing several alveoli to merge, reducing the respiratory area of the lungs. Chronic bronchitis is a long-term inflammation of the bronchial tubes, causing mucus-accumulation and breathing difficulty (5).

2.1.1 Causes

The single most common cause of COPD is inhalation of pollutants, e.g. cigarette smoke (5).

2.1.2 Symptoms

Initially, symptoms of COPD are often mild or non-existent. As the disease develops, symptoms often are (2,5):

➢ Persistent, mucus-producing coughing
➢ Shortness of breath
➢ Chest tightness
➢ Wheezing
➢ Cachexia and Sarcopenia

2.1.2.1 Cachexia and Sarcopenia

Two frequently occurring symptoms of developed COPD is cachexia and sarcopenia.

Cachexia has been characterized by an overall severe loss of both muscle and fat mass and increased metabolic activity. Additionally, the underlying disease can disrupt the signal balances of appetite stimulation, resulting in anorexia. The diagnosis criteria of cachexia are an unintentional weight loss of more than 5 % in six months and a fat free mass index (FFMI) below 15 kg*m⁻² for women and 17 kg*m⁻² for men. Cachexia-patients, if untreated, risk developing sarcopenia as well (6).

Sarcopenia is a muscle wasting syndrome, often associated with aging, but also prominent in COPD. As sarcopenia develops, respiratory function is impaired, which can have devastating results in COPD patients already struggling with breathing. Diagnosis of sarcopenia has been characterized by two criteria (2,6):

1. A skeletal muscle mass index (SMI) equal to or below the SMI mean minus two standard deviations of healthy people of the same ethnicity and sex, aged 20 to 30 years. SMI is defined as the lean appendicular mass divided by the person’s height squared.
2. A walking speed below 0.8 m/s in a 4m walking test.

The loss of muscle can lead to a tissue replacement, in which fat is accumulated in place of the wasting muscle, causing sarcopenic obesity, also known as hidden obesity (2).

2.1.3 Diagnosis - spirometry

A common method for diagnosing COPD and to track respiratory capability over time, is spirometry. Spirometry is designed to assess lung functionality and consists of the patient exhaling at into a tube connected to a machine, the
spirometer. Two key measurements are produced, forced vital capacity (FVC) and forced expiratory volume (FEV1). FVC specifies the total amount of air the patient can exhale forcefully, after taking an as deep breath as possible. FEV1 is the volume of air the patient can exhale forcefully from the lungs in one second. A low FVC is indicative of restricted breathing, while a low FEV1 is indicative of more severe respiratory obstruction (7).

The cutoff limit of lung function to consider a COPD diagnosis is set either through a fixed ratio (FEV1 / FVC < 0.7), or a lower limit of normal FEV1 / FVC ratio defined by the lower fifth percentile of a reference population (8). The two methods result in different cutoff values, and COPD-related hospital admissions have been shown to be higher in the intermediate population than in populations with normal lung function. The higher threshold, which is the lower limit of normal, may as a result wrongly declare a patient as healthy (8). The Global initiative for Chronic Obstructive Lung Disease (GOLD) favors the fixed FEV1 / FVC ratio (9).

Once a subject has been confirmed to have a FEV1 / FVC ratio below the 0.7 threshold, the severity of obstruction can be classified by the GOLD airflow limitation severity table (9):

<table>
<thead>
<tr>
<th>Stage</th>
<th>FEV1/FVC% of predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>≥ 80%</td>
</tr>
<tr>
<td>2</td>
<td>50% ≤ FEV1 &lt; 80%</td>
</tr>
<tr>
<td>3</td>
<td>30% ≤ FEV1 &lt; 50%</td>
</tr>
<tr>
<td>4</td>
<td>FEV1 &lt; 30%</td>
</tr>
</tbody>
</table>

The predicted value for FEV1 used in the GOLD stages in men can be calculated using the formula (10):

\[ 4.30 \times \text{height} - 0.029 \times \text{age} - 2.49 \]

For women:

\[ 3.95 \times \text{height} - 0.025 \times \text{age} - 2.60 \]

This formula does not take patient ethnicity in account, which has been proven to affect lung size and as a result spirometric performance (10). Also, while height certainly affects lung size, inaccuracy due to the confounding effect of osteoporosis should be considered, since it may result in a slightly lower predicted FEV1 value. Osteoporosis is a common symptom of COPD, and should be accounted for when comparing a predicted and measured FEV1 (10).

### 2.1.4 Previously proposed phenotypes

In 1968, two phenotypes for end-stage COPD was defined based on the emphysema and bronchitis conditions. In later years an additional categorization has been proposed, which is based not only on the origin of obstruction, but also on metabolic impact. The three phenotypes presented and their specific characteristics are (2):

1. **Cachexic and emphysema**, characterized by:
   a. Skeletal muscle mass and fat mass loss
   b. Muscle fiber atrophy
   c. A shift from muscle fiber type 1 to 2, causing a decreased muscle function
   d. Weakened bone structure (osteoporosis).

2. **Obese with chronic bronchitis**, characterized by:
   a. Increased subcutaneous and visceral adipose tissue
   b. Arterial stiffness and increased cardiovascular risk

3. **Sarcopenic with hidden obesity**, characterized by:
   a. Loss of skeletal muscle mass
b. Muscle fiber atrophy

c. A shift from muscle fiber type 1 to 2, causing a decreased muscle function

d. Preserved but redistributed fat mass, increasing visceral adipose tissue, arterial stiffness and an increased cardiovascular risk

COPD is a complex disease, where interventions can give drastically different responses depending on the patient and nature of that individuals’ condition. Different facets of an individual’s disease may require different intervention approaches (2).

It is important to question the validity of using phenotypes to categorize COPD patients at all, especially using the original definitions of “either bronchitis or emphysema”. GOLD is even claiming that emphysema as a term is often used clinically incorrectly, and that chronic bronchitis by their definition only rarely occurs in COPD patients (9).

2.2 BMI AND ANTHROPOMETRIC SURROGATE PARAMETERS

Currently, health evaluation is often estimated through the Body Mass Index (BMI). It is also used to divide people into categories, supposedly grouping them with people of similar body composition and disposition to develop metabolic diseases (11).

While BMI sufficiently correlates with body fat and future health risks on a population scale, the limitations of the parameter quickly become prominent on an individual level. For example, BMI only incorporates weight and height, while ignoring other factors such as fat distribution and muscle mass.

BMI is a surrogate parameter for either being underweight, overweight or neither. BMI does not, as mentioned, define where potential excess fat is located. This is problematic because specific fat distributions have been significantly linked to metabolic complications, while other distributions actually seem to decrease susceptibility (12).

Additional anthropometric parameters are generally needed to approximate fat distribution, such as waist circumference. However, waist circumference combined with BMI values are less accurate when measuring women than men, and additionally only aims to estimate visceral adipose tissue (13). Based on the mentioned observations it would seem that single-parameter approximation methods perform poorly in precision medicine.

2.3 BODY COMPOSITION PROFILING

In disease risk profiling, certain metabolic phenotypes are considered to be more likely than others to develop diseases. Phenotypes are discrete setups of metabolic properties into which different patients are divided (2). The issue, however, is that the human metabolic diversity exists on a spectrum rather than a rigid set of discrete phenotypes. This means that oftentimes part of an individuals’ metabolic profile matches to one phenotype, while other parts match to others. To accurately predict which health risks an individual pertains, an individual phenotype would be needed for everyone.

2.3.1 The virtual control group

Using large data bases, a multidimensional network of a selected group, or all subjects available, can be generated. The network dimensions represent different, normalized, data variables. The scanned individual is added into this network based on the data obtained through the scan and possibly additional, external parameters. The individual will be clustered with other subjects, whose data are similar to their own. The closest matches (a number usually specified around 50-100, with disease case prevalence limits as an optional requirement for disease analysis) are used as a virtual control group (VCG) for the patient.
The VCG will be gender-specific. No male subjects will be present in a females’ VCG, and vice versa.

The VCG is based on matching variables specified by the user. If a VCG is created using VATi and ASATi as matching variables, a scatter plot of the patient and the VCG would place the patient in the middle of a VCG “circle” (Figure 1). When the network has been set up, external data such as prevalence of diseases related to metabolic phenotypes can be scanned for.

2.3.2 Propensity

The propensity calculation of an individual to develop a selected disease is based on the share of case subjects in the VCG. By setting a requirement, in addition to the minimal size of the VCG, you can choose to include a minimum number of cases. The cases are represented by a binary column.

For example: If in the data there is a case-column for Type-2-Diabetes, which is 1 if the patient has diabetes and 0 if not. By choosing this column as the “case-column”, and setting the minimum case number to 20, the VCG for a subject will contain at least 20 subjects with a 1 in the Type 2 diabetes column. The VCG will continue to grow until this requirement is met. The propensity (ppDisease) of that patient is then calculated using the following formula:

\[
ppDisease = \frac{\text{Number of cases}}{\text{Total VCG size}}
\]

The continuous growth until a prevalence threshold is met is performed to avoid two individuals receiving identical propensities for a condition, which would risk them becoming interchangeable in sorting. Also, in a control group with too few cases there can be no valid statistical tests performed; the dynamic growth aims to avoid that issue.

Figure 1: A VCG of an individual created using matching variables VATi and ASATi. In a scatter plot of the matching variables, the VCG creates a circle around the patient, if the distribution of subjects is even. The patient ID is hidden.

ASATi = Abdominal Subcutaneous Adipose Tissue; VATi = Visceral Adipose Tissue
2.3.3 Body composition profile

In the field of acquiring accurate multi-parameter body composition data, magnetic resonance imaging (MRI) is currently the gold standard, and can provide accurate and direct quantification of body composition parameters such as skeletal muscle mass, body fat percentage and body fat distribution (14). An MRI image example can be seen in Figure 2 (15).

Through resonance differences in water and fat, different tissues can be distinguished and quantified, through methods such as the Dixon method (16).

By utilizing MRI images, information regarding body composition of the scanned individual is obtained. First, inhomogeneity in intensity of images is corrected, using pure adipose tissue as an internal signal reference. The fat and water images are then merged, producing a composite image set, covering the area between the neck and knees. Non-rigid ground truth atlas-based registration is then used to categorize the acquired volumes. The atlases used have been selected using histograms of biomarkers and visual inspection of their MRI images. A voxel, a three-dimensional pixel, is assigned a label if more than five atlases agree which label the voxel corresponds to (14).

The following measurements used in this thesis are obtained directly through the MRI protocol (17):

- **Visceral Adipose Tissue (VAT)**. VAT has been defined as adipose tissue within the abdominal cavity, excluding all adipose tissue and lipids outside of the abdominal skeletal muscles, within and posterior of the spine as well as posterior of the back muscles (14).
- **Abdominal Subcutaneous Adipose Tissue (ASAT)**. ASAT has been defined as subcutaneous adipose tissue in the abdomen, starting at the top of the femoral head reaching to the top of the thoracic vertebrae T9 (14).
- **Intramuscular adipose tissue (IMAT)**. IMAT is assessed based on water/fat images. Normal muscle is expected to have a fat value of 0, while muscle sections approaching or exceeding 50 % fat is considered as fat infiltrated (18).
- **Liver Proton-Density Fat Fraction (lffp0)**. The liver signal is divided into its water (W) and fat (F) signals. In simplified terms the fat fraction, n, is calculated using the two values (19):
  \[ n = \frac{F}{W + F} \]

- **Lean Thigh Muscle Volume**. The thigh muscle volume include muscles gluteus, iliacus, adductor and hamstring (collectively the posterior thigh muscles) as well as quadriceps femoris and sartorius (collectively the anterior thigh muscles (14). The measurements result in the variables rulf (right upper leg front), lulf (left upper leg front), rulb (right upper leg back) and lulb (left upper leg back).

Using these data, a body composition profile (BCP) specific to the individual being scanned is created. The standardized visualization of the BCP is referred to as the BCP star (Figure 3) and is a radar chart with axes depicting the patient, VCG and metabolically disease free (MDF) values for six variables. The variables used are:
➢ IMAT
➢ Lff10p
➢ VATi. When used as an index VAT is divided by the individual’s height² and is referred to as VATi
➢ Abdominal Total Adipose Tissue index (ATATi). ATAT is the summation of VAT and ASAT. When used as an index ATAT is divided by the individual’s height² and is referred to as ATATi.
➢ Weight-to-muscle ratio (MR). The Weight-to-muscle ratio is calculated by dividing the individuals’ total weight by the thigh muscle mass (lulb, rulb, lulf and rulf).
➢ Fat Ratio (FR). An assessment of an individual’s ability to carry their fat. Calculated through:

\[
\frac{ASAT + VAT}{ASAT + VAT + lulb + rulb + lulf + rulf}
\]

These BCP variables differ in both distribution and magnitude, so to fit on the same radar chart their values are mapped through a logarithmic sigmoid transfer function, resulting in values for all variables between 0 and 1, where 0 is in the center of the radar chart and 1 is at the end of the spokes. The characteristic star shape of the MDF reference was created by mapping the MDF values to fixed distances on the six spokes. The ectopic fat variables (VATi, IMAT and lff10p) have their reference values mapped to 0.15, while the other three variables (ATATi, MR and FR) are mapped to 0.6 (3).

When plotting the BCP star, the patient values are connected and displayed through a red line. The MDF is represented by blue dashed lines and the VCG consist of a shaded field, displaying the interquartile range of the group.

2.4 SOFTWARE

2.4.1 R

R is an open source programming language and environment for statistical analysis and visualization. AMRA has used R to develop most statistical methods upon which this master’s thesis is based.

2.4.2 Mathematica and the Wolfram Language

The Wolfram Language (WL) is an extensively documented multiple-paradigm programming language developed by Wolfram Research. The language has fluent custom interface construction properties, allowing for quick production of data interaction and visualization. Mathematica is a computing system in which the Wolfram Language is used.
2.4.3 **RLink**
There has been compatibility developed for R and Mathematica through a command called RLink. Through this, R commands can be executed within the Mathematica framework, and output can be interpreted as Wolfram Language. This functionality enables previous work in R to be seamlessly implemented into the Mathematica framework.

2.5 **GRAPHICAL CONSIDERATIONS**
The tool resulting from this project will be presenting graphical visualization of patient, VCG and reference population data. When presenting a reader visual representation of data, there are various aspects of visualization to consider (20).

2.5.1 **Graphical excellence**
Edward Tufte, esteemed statistician and artist, states that to achieve excellence in statistical graphics, complex ideas should be conveyed with clarity in an efficient and precise manner. The graphical display should, for instance (20):

- Show the data.
- Induce substance analysis rather than analysis of the underlying methodology or technology.
- Encourage subconscious detection of differences in the data.
- Serve a reasonably clear purpose.
- Give the greatest number of ideas in the shortest time with minimum ink and visualization area.

2.5.2 **Graphical integrity**
The data displayed through graphical means must be reliable. In order to achieve this, the following points should be considered (20):

- The surface area occupied should be proportional to the numerical quantity represented.
- To avoid graphical distortion and ambiguity, clear and detailed labelling should be used. Important events should be highlighted.
- Data is the only item that should vary in a visual data representation, design should be constant.
- The number of dimensions used to visualize data should be equal to the dimensions of the data.
- Graphics should display the whole picture, and not take data out of context.

2.5.3 **The problem with physically multi-dimensional plots**
When a multi-dimensional data set is to be visualized, attempts are often made to incorporate more dimensions into a single plot. 3D scatter plots or density plots can be created and even supplemented with a fourth dimension through the color spectrum and a fifth dimension through point size. The result is indeed a plot of multivariate information, but to understand it takes significant effort and requires interactivity, reducing compatibility with exporting the graphics to physical media or other platforms. One must ask whether it would be preferable to replace the multi-dimensional graphic with two or three bi-variate graphics. Though the idea of coloring in an additional dimension is an intriguing one, our minds are do not see color automatically as a spectrum, and can for example have a hard time distinguishing one shade of blue from another (20,21).

2.5.4 **Multifunctionality of graphics**
Graphics can be viewed at different depths (20):

- A superficial layer which can be seen at a distance, where the eye catches patterns of the underlying data.
- The fine structure of the data, viewable up close.
- Implicit details, shining through by reading between the lines of the graphic.
2.5.5 Data density
The human eye is capable of distinguishing details at a small level. For example, a map of all 30000 communes in France can be studied without much difficulty in a 175 cm² area. Because of this, if an issue arises where the amount of graphics needed to convey a message out-sizes the boundaries of the general computer screen or presentation slide, a reduction in graphics size is not unreasonable (20). Labels and other text should still be readable, however, and a consideration must be made as to at which level of detail an observer is expected to view the data.

2.5.6 Other considerations
In addition to the above mentioned graphical considerations, the following should be considered (20).

➢ Serif font should be used for facilitated reading for longer texts.
➢ Abbreviations should be minimal if used at all.
➢ Unnecessary coloring should be avoided in graphics.
➢ Graphics should stretch further horizontally than vertically, due to our natural practice in noticing deviations from the horizon. The cause should be horizontal and the effect vertical. The Golden Ratio setting is the default in most plots in the WL (Figure 4).

2.5.7 Plot types considered
Mathematica has a range of visualization alternatives to offer, several of which aim to visualize comparisons of vectorized data.

2.5.7.1 Smooth Histogram
The SmoothHistogram[] command accepts one or more vectors and by default plots the probability density function of the vector(s) (22). Several data sets can be entered. Figure 4 shows an example of this plot.

2.5.7.2 Sector Chart
The SectorChart[] command accepts bi-parametric input arrays, and creates a pie chart in which each sector also expands outwards depending on the second input-value. Locking the first input value could provide an alternative to the traditional and critiqued (20), pie chart, while retaining the single-parametric relational comparison of the classic pie chart. In Figure 5, sectorchart b and c both display the same data, but the magnitude of the differences is easier to spot in sectorchart b.

Figure 4: Plot of the probability density function for the VAT-vectors of the female UKBB-subjects and the VCG of a female COPD patient. VAT = Visceral Adipose Tissue; VCG = Virtual Control Group

Figure 5: a: SectorChart of a bi-parametric data set with both inputs containing data. b: uni-parametric data set with the first input locked to 1. c: uni-parametric dataset with the second input locked to 1. The data used in this image is dummy data.

Figure 6: 4D smooth density plot created by ListDensityPlot3D.
2.5.7.3 ListDensityPlot3D
ListDensityPlot3D accepts quad-parametric input arrays and thus allows for analysis of 4 parameters simultaneously, in which the fourth is represented by color (Figure 6). The plot is a density plot, and linearly interpolates values to give color changes.

2.5.7.4 Smooth Density Histogram
The SmoothDensityHistogram[] command accepts bi-parametric input arrays and by default produces a 2D probability density heat map of the two parameters (22).

In Figure 7 an example of the SmoothDensityHistogram is shown over the age and BMI data for subjects.

2.5.7.5 3D Scatter Plot
ListPointPlot3D[] accepts tri-parametric input array and creates a 3D scatter plot. This allows for analysis of three parameters simultaneously. Several datasets can be entered simultaneously, as seen in Figure 8.

![Figure 7](image1.png)

Figure 7: Probability density plot for age (horizontal axis) and BMI (vertical axis) vectors of the female UKBB population.

UKBB = UK Biobank

![Figure 8](image2.png)

Figure 8: A ListPointPlot3D of the ASATi, VATi and IMAT vectors of the male UKBB data set (grey), the VCG of a male subject (red) and the subject (blue). In Mathematica, the 3D space can be moved freely around all axes.

ASATi = Abdominal Subcutaneous Adipose Tissue; VATi = Visceral Adipose Tissue; IMAT = Intramuscular Adipose Tissue; UKBB = UK Biobank; VCG = Virtual Control Group
3 Method

3.1 Method Overview

The methodology used in this project is summarized in Figure 9. The existing R code and data was used to generate the first VCG. Once the basic algorithm was set up, the resulting data from the VCG could be used to develop the first visualizations. This eventually allowed for COPD specific analyses, which through iterative development produced more COPD specific visualization methods. Once visualization of VCG-based COPD assessment had been established, the data could be screened for intrinsic metabolic differences within COPD. To evaluate the tool

Figure 9: Work process overview. The arrows indicate that the arrow source activity was necessary to complete before starting the arrow destination activity. The colors separate the process into general activities, data visualization and VCG-based diagnostics.

WL = Wolfram Language; VCG = Virtual Control Group; srCOPD = Self-reported COPD; ppXX = propensity for XX.
and its COPD assessing capabilities, the algorithm and system was tested on a select number of individuals from the UK Biobank data set.

3.2 IMPLEMENTATION OF R INTO WL
The first step of this master’s thesis was to transfer code from R into WL. RLink was used intermittently with translation of R commands into corresponding WL commands. If R was considered more suitable to perform the desired task, then RLink was used. If extensive interaction of a variable or dataset was needed, a transition into WL was performed.

3.3 THE DATA
The main data source used in this project was the UK Biobank. A smaller study that focused on a group of COPD subjects was also available.

3.3.1 UK Biobank
There are several databases containing data from thousands of subjects; both anthropometric as well as clinical data is readily available. One such databank is the UK Biobank (UKBB). This international resource, based in the United Kingdom, contains data from 500 000 people who at the project’s inception years 2006-2010 were between 40-69 years of age. The data available for each anonymized individual has been acquired through measures such as MRI and spirometry, samples and questionnaires (23).

The UKBB contains thousands of data columns for each subject, although not all of them contain a value. The columns used in this project are listed in Table 3.

A vital additional layer to the UKBB data is their imaging study. In 2006 a project was set up with the goal of scanning 100 000 people using MRI, with actual scans starting in 2016.

In this project, datasets including MRI data of 6021 individuals was be available, potentially with additional subjects becoming available at a later date (23).

The data set used originating from the UKBB in this project consisted of 6021 subjects. This dataset was referred to as the UKBB data set.
Table 3: Columns from the UKBB data set. The columns hcb_trunc15, ppCHD, ppT2D and ppHCB have been calculated using other columns in the UKBB data set.

<table>
<thead>
<tr>
<th>Column name</th>
<th>Description</th>
<th>Column name</th>
<th>Description</th>
<th>Column name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>aid</td>
<td>AMRA ID</td>
<td>gender</td>
<td>Subject gender</td>
<td>lff10p</td>
<td>Subject lff10p</td>
</tr>
<tr>
<td>eid</td>
<td>Electronic ID</td>
<td>bmi</td>
<td>Subject BMI</td>
<td>asati</td>
<td>Subject ASATi</td>
</tr>
<tr>
<td>age</td>
<td>Subject age</td>
<td>vat</td>
<td>Subject VAT</td>
<td>atati</td>
<td>Subject ATATi</td>
</tr>
<tr>
<td>height</td>
<td>Subject height</td>
<td>asat</td>
<td>Subject ASAT</td>
<td>imat</td>
<td>Subject IMAT</td>
</tr>
<tr>
<td>weight</td>
<td>Subject weight</td>
<td>FR</td>
<td>Subject FR</td>
<td>vati</td>
<td>Subject VATi</td>
</tr>
<tr>
<td>MR</td>
<td></td>
<td>VATr</td>
<td>VAT/(VAT+ASAT)</td>
<td>ttvi</td>
<td>Subject TTVi</td>
</tr>
<tr>
<td>hcb</td>
<td>Health Care burden (see Table 1)</td>
<td>Hcb_trunc15</td>
<td>Hcb truncated to 30 for the last 15 year prior to scan</td>
<td>fvc</td>
<td>Subject FVC</td>
</tr>
<tr>
<td>ppCHD</td>
<td>Subject propensity for CHD</td>
<td>ppT2D</td>
<td>Subject propensity for T2D</td>
<td>ppHCB</td>
<td>Subject propensity for HCB</td>
</tr>
<tr>
<td>Fev1</td>
<td>Subject FEV1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3.2 The COPD-D study
In 2014 a study was conducted with the purpose of analyzing the effect a deficiency of vitamin D has on COPD patients. The study resulted in a dataset of 9 women and 6 men with mainly spirometry and MRI body composition data. The study was being updated in early 2018 and could have resulted in additional information during the course of this project. Currently, lff10p data is not available in the COPD study.

The data available from the vitamin D deficiency study contains information from 15 patients with severe COPD. This data was referred to as the COPD study and could be used as a validation tool in COPD identification within the UKBB data set.

3.3.3 Estimating COPD datasets in the UKBB
No extensively filled data column was available in the UKBB data set containing information of presence or absence of diagnosed COPD. FEV1 and FVC values were available however, which allowed for an estimation of COPD presence using the FEV1/FVC diagnostic. A subject with a FEV1/FVC<0.7 was assumed to have COPD. This estimation results in a COPD prevalence of 14.2 % (406 out of 2861) in the male population and 11.3 % (356 out of 3156) in the female population of the UKBB data set. The estimated global prevalence of COPD was 11.7 % in 2010, with a overrepresentation of men as compared to women (9). Since this single metric is not a sufficient diagnostic tool, this data set was referred to as the “soft” COPD data set.

One sparsely filled column was available stating the age at which the subject was diagnosed with either emphysema or bronchitis (the main types of COPD). 54 women and 55 men had entered a value into this column, which creates an additional data set to use as a robustness test when attempting to discover COPD. This data set was referred to as the self-reported COPD (srCOPD) data set.

Two columns contained and measurement-based diagnoses of sarcopenia, one through MRI and one through DXA. A new column was created in which the subject was given the value 1 if they had a 1 in either the MRI or DXA
columns. Through this column, the sarcopenia data set was formed. This data set contained 671 men and 916 women. The diagnosis for sarcopenia for these methods is a threshold of the total thigh muscle volume index (ttvi). The “mdf” column represents subjects deemed metabolically disease free. This column was used to generate the mdf data set, which was used to generate healthy VCGs to estimate individual reference values. To clean this data set from potential COPD members, all soft COPD members present in the MDF data set were removed. The MDF data set contained 833 women and 693 men after being cleaned.

3.4 CREATING THE VIRTUAL CONTROL GROUP

3.4.1 Identification
The original script for creating a VCG was specific to the UKBB data and as such utilized an ID column in that database which was absent in the COPD study data. Fortunately, another ID system was available, aID, which existed in both data sets. To facilitate interchanging of analyzed data sets, all references to the ID of a patient or a member of their VCG was changed to a generic ID-reference, which is specified at the start of the VCG script. In this project, that ID is always set to aID but the code is made to run with any ID system shared by both data sets.

3.4.2 Data columns
Creating a virtual control group required two data sets, one with subjects who would be given a VCG, and one from which the VCG members would be generated. To identify which data was available in both data sets, an algorithm was created to compare the heads of all columns in the subject data set to the VCG source data set. The algorithm generated a button for each of the common columns to allow the user to select which of the common columns were of analytical interest and should be transferred into the VCG once it had been completed (Figure 10).

![Figure 10: Buttons for which columns to transfer to the VCG. A pressed (blue) button means that the respective column will be transferred.](image)

Some columns containing the same data had different names in the UKBB data set as compared to the COPD study. Other column names contained dots, which produced errors when automatically creating parameter names based on column names. Because of this, columns containing the same information but under different names had their names adjusted to match each other, while columns with dots had the dots removed.

3.4.3 Matching variables
Depending on the purpose of analysis, different VCGs could be created for a single individual, either by using different matching variables or different data sets from which to create the VCG. The matching variable input had to be case-identical to the column names, e.g writing “ASATi” would not be interpreted as the column “asati”.

3.4.4 The output
Once the VCG had been created using the matching variables, all data columns selected for all members of the VCG and the source data set were saved into a .csv file. An additional file was also created, containing additional data for all chosen variables, such as the subjects’ effect size when compared the values of the VCG.

The members of the source data set that were not selected for the VCG are referred to as the reference population.
3.5 DEVELOPING VISUALIZATION

Metabolic diseases are complex and require a wide understanding of a multitude of patient specific variable values and ratios. To provide wholesome characteristics of a patient, more than one visualization was used. The goal was to see the metabolic properties of an individual and be able to conclude whether these were abnormal for that specific individual. Another goal was to put the individual in the context of COPD and assess the metabolic COPD tendencies the individual expressed.

3.5.1 The Patientsector

The SectorChart[] command was identified early in the project as a possible multi-factorial visualization of an individuals’ values, and went through 6 iterations. Screen captures of these iterations are available in Figure 12.

3.5.1.1 Iteration 1

The first SectorChart iteration was a highly hands-on, non-automated, two-layered plot of specific variables hard coded into the script that generated the plot. It compared the values of the individual, his or her VCG and a reference population of the same gender. To put the values in relation to each other all variables were divided by the mean of the reference populations’ counterpart. Three plots were created for the reference population, the VCG and the patient. For the reference population and the VCG, the first row showed the mean, the second showed the maximum and the third the minimum values (all divided by the reference populations’ corresponding mean). The inner two variables were FEV₁ and FVC, because they were considered to be of particular interest. The three patient plots were identical and were duplicated as such to provide easier reference to the other groups’ values.

3.5.1.2 Iteration 2

The next iteration of the SectorChart was similar, but with one dimension instead of two. This was done to more easily distinguish abnormal values. The variable names were also added to allow the user to identify each variable. The plot variables were still hard coded.

3.5.1.3 Iteration 3

A more automated approach was attempted, in which a selection of plot variables in the same button-form as the creation of the reference data was used. The automation was performed to provide the user with the ability of choosing which variables were of interest, as well as to provide the opportunity to use data sets with other variables than those used for this project.

All sector charts were set to divided by the median of either the reference population or the VCG, or the patient’s own values. The switch from mean to median was performed since not all variables had a normal distribution. The second and third iterations were identical in appearance.

3.5.1.4 Iteration 4, 5 and 6

The fourth iteration, now called the Patientsector, was a single SectorPlot of the patients’ values divided by the VCG median, plotted over a circle representing the VCG median over itself. This iteration was intended to be viewed with additional plots highlighting other aspects of the data. A problem with this approach was that if the patient values were greater than the VCG median, the circle could not be seen. An alternative, iteration 5, displays the circle on top of the sectors instead. Finally, to prevent confusion regarding the purpose of the coloring, all sectors were set to one single color, since their presence was only aesthetic.

3.5.2 The density histogram

The density histogram had been used previously when performing VCG analysis, and as such had already been processed through iterative improvement efforts. Still, some development went into improving the data-ink ratio.

The first and most significant difference is the implementation from R into WL. Much of the VCG algorithm is performed through the RLink, but this visualization was completely translated into WL for facilitated graphical
adjustments. In R, the plots were hard coded for specific variables, and these were then saved as several .pdf files into a specified folder. Instead, the data required for the plot is now saved as a .csv file, and the plot is generated upon command for the studied individual. The selection of which variables to plot is done effortlessly through the auto-generated selection menu.

### 3.5.2.1 Iteration 1
The first WL iteration is similar to the original R version, but the background grid was removed for a slimmer appearance. The outline and fill color has been changed to the same color, and the colors used are changed as well to the WL default theme.

### 3.5.2.2 Iteration 2
To improve readability in printed non-colored format, the plot line and fill opacity was changed for the reference population, producing a higher contrast from the VCG. The effect size of the individual as compared to the VCG (VCGsd) and the reference population (GPsd) is printed in the upper right corner, with a comparison of the two through subtraction of the absolute values of GPsd from VCGsd. A positive difference indicated that the patient differs more from the VCG than the reference population, which might indicate abnormal values for that variable.

### 3.5.3 Creating and importing the BCP star
The plotting of the BCP star was performed in Python and the plot was saved as a .pdf file. Through Mathematica, the terminal command for running the script was called through the Run[] function. The BCP plot is created together with the VCG and is available for viewing with other patient analyses.

### 3.5.4 The GOLD stages
If an individuals’ FEV$_1$/FVC value falls below 0.7 then they can be placed in one of the four stages of airway obstruction, as defined by GOLD (see Table 2). Since the stage spaces are defined by the individuals’ own values, a visualization of the stages on a continuous spectrum was deemed to facilitate interpretation of airway obstruction severity. To achieve this, a number line plot was used, which is a single dimension, horizontal visualization tool. Utilizing verticality, the different stage intervals were separated by height to facilitate readability.

### 3.5.5 The 3D VCG exploration (VCG-X)
The 3D sector plot allows for analysis of the VCG together with the patient and the reference population in a 3D space. By adding manipulation abilities through variable selection and plot-range sliders, it becomes a powerful tool with which to assess the patient in all scale-based measurable aspects. The final version was referred to as the VCG-X. While the manipulation tools were added incrementally, the VCG-X did not change much in appearance during the project.

### 3.5.6 Disease prevalence
To provide information of symptoms other than those directly related to COPD, the VCGs of each subject is scanned for presence of Type-2-diabetes, coronary heart disease and a hcb_trunc15 value above 0. The subject’s VCG percentage of these diseases are displayed together with their prevalence in the UKBB data set and an odds ratio depicting the significance of the difference, corrected for sample size. Together with these, the VCG prevalence of srCOPD and sarcopenia were also displayed.

### 3.6 VCG-based COPD diagnostics
In order to visualize an individual’s COPD characteristics using VCGs, the capability of assessing COPD using the available data had to be analyzed.

Propensity values were calculated for presence of VCG members belonging to the srCOPD data set. Several different matching variable sets were tested, starting with the least ectopic BCP variables ASATi, FR and MR. The
performance of the propensity was evaluated through receiver operating characteristics (ROC) curves and area under the ROC (AUROC) values when set to diagnose self-reported COPD. The propensity value was given the abbreviation “ppsrCOPD”.

Propensities were also calculated or obtained through previous work for T2D, CHD, HCB and sarcopenia.

3.6.1 Virtual Lung Obstruction
To increase the data driven nature of a predicted FEV$_1$ value, all members of the UKBB data set were given a VCG created only from UKBB members in the MDF data set with FEV$_1$/FVC>0.7. This subset of the MDF data set was named “extrememdf.csv”. The matching variables were the non-ectopic BCP variables ATATi, MR and FR.

The avoidance of ectopic variables and cases aimed to predict a healthy FEV$_1$ VCG-value for a BCP that was as similar to the individual as possible. The predicted FEV$_1$ was taken as the median of the healthy VCG, FEV$_1$VCGmed.

Level of obstruction was estimated for each subject in the UKBB data set by FEV$_1$VCGmed/FEV$_1$ and was referred to as the virtual lung obstruction, or VLO$_{med}$. 4 subjects had extremely low FEV$_1$ values which made their virtual lung obstruction value several times higher than all other UKBB members. These subjects were assumed to have been incorrectly measured and were removed in subsequent tests.

3.6.1 Assessing diagnosis performance
To determine the diagnostic performance of ppsrCOPD and VLO$_{med}$ in comparison to the traditional GOLD scale, Pearson correlation tests were performed on virtual lung obstruction and ppsrCOPD against hcb_trunc15 and srCOPD. ROC analysis was also performed to assess the VLO$_{med}$ ability to identify srCOPD.

To more easily compare the two predicted values, the GOLD scale was translated into a continuous variable, “GOLDcontinuous”, by removing the cutoffs and instead use the FEV$_1$predicted/FEV$_1$ formula. Similarly, a propensity value for hcb_trunc15 was used in order to achieve a continuous variable (ppHCB).

Hcb_trunc15 was deemed an appropriate health indication, since it’s value represents the subjects total number of nights at a hospital during the last 15 years, truncated to 30 times, corrected for nights related to pregnancy.

3.7 Exploring COPD phenotypes
In order to stratify a potential COPD diagnosis using metabolic measurements, the srCOPD cases were analyzed through BCP characteristics and propensities for other conditions.

All members of the srCOPD data set had their BCP stars plotted in a table in order to detect signs of different phenotypes within the group. These potential phenotypes were then studied to explore their propensities for other conditions.

To determine if the potential BCP characteristics could be found through reverse-engineering of high propensity data sets, the 10% UKBB data set members for each gender with the highest ppsrCOPD were extracted and further divided into subgroups with the lowest 10% of the propensities sarcopenia, T2D, CHD and HCB respectively. These groups in turn had their BCPs analyzed and were put in scatter plots for the other propensities in order to further stratify them.
RESULT

This section presents, in order, the visualizations for the patient's metabolic and spirometric condition as compared to both the VCG and the general population; the algorithm for generating the visualization; the VCG based COPD diagnostics; the tool resulting from the visualization and algorithm; usage of the tool to characterize COPD and finally an example where an individual is analyzed using the tool.

4 DEVELOPING VISUALIZATIONS

4.1 The Patientsector

The six iterations of the Patientsector can be viewed in Figure 13 and Figure 13. The second and third iterations were identical in appearance, and are both represented in Figure 13 B.

Figure 12 A: First iteration of the SectorPlot. In the upper row the values of the general population, VCG and individual are divided by the general population mean. The next two rows of sectors are divided by the general population maximum and minimum respectively. The inner circle represented the FEV1 and FVC values, while the surrounding were BCP and blood values. No variable labelling was used, creating a difficult or even impossible to understand plot.

Figure 12 B: The third, automated, iteration of the sector plot. Here the top three plots are divided by the general population median, the second row by the VCG median and the third row by the patients’ values. The variables displayed are chosen by the user.
4.1.2 The density histogram

The multi-variate Patient sector is complemented by the variable-specific density histograms. The density histograms provide information of how the VCG is different from the reference population, and how the patient compares to the two. Adding the effect size of patient-VCG and patient-reference population, allows for a statistical assessment; is the patient value more like the VCG or is the patient more like the reference population? The development of the density histogram can be studied in Figure 14.
Figure 14: a) Example of the “plot1”.pdf for one patient’s density histograms, as created originally in R. Three other .pdf files were created for each patient.

b) The first WL iteration of the density histogram. Here, the age distributions are displayed, with the patient being younger than the median of both VCG and reference population.

c) Final version of the Density Histogram. The filling opacity and plot line style of the reference population is what was changed after the R to WL translation, together with the standard deviation comparison.
4.1.3 The importing of the BCP
The BCP is imported as a .pdf into the Mathematica framework. When importing the .pdf file the transparency of the VCG is lost, sometimes resulting in difficult to read overlaps of VCG, MDF and the individuals’ values (Figure 15).

4.1.4 The GOLD scale
The GOLD scale is a single-variable visualization. The dimension visualized is FEV1. The four stage intervals were given separate vectors and were plotted in a stair-like manner, as to utilize the vertical spectrum to indicate obstruction severity and clearly separate the stages.

ASATi = Abdominal Subcutaneous Adipose Tissue; VATi = Visceral Adipose Tissue; IMAT = Intramuscular Adipose Tissue; MR = Muscle Ratio; lff10p = Liver Proton Density Fat Fraction (10 points); FR = Fat Ratio

GOLD scale

Figure 15: **Left**: BCP star as it appears when imported into Wolfram Mathematica. **Right**: The BCP star as it appears when using a .pdf reader.

Figure 16: The GOLD scale visualization. The vertical line places the individual on their individual spectrum indicating airway obstruction level. **FEV1** = Forced Expiratory Volume first second; **GOLD** = Global initiative for Obstructive Lung Disease.
4.1.5 **The VCG-X**

The VCG-X allows the user to compare a variable of the individual in a 3D space to any other 2 variables and their relation to the VCG and reference population. The variables displayed can be changed by drop-down menus. Clicking “Update” will rerun the plotting within the VCG-X. The plotted data sets are given in the upper right corner and will remain in place regardless of the rotation within the 3D-space.

4.2 **The IICE Algorithm**

The Interactive Individual COPD Evaluation (IICE) tool is dependent on the IICE algorithm, which is responsible for the generation of the VCG and all analysis graphics associated with the patient analysis. The IICE algorithm was written within the Mathematica framework, and works within the confines of the Mathematica “cells”. These cells each contain one section of the algorithm and can be executed one by one, or all at once by evaluating all cells tagged of a certain “tag”. The cells need to be evaluated in the correct order (top to bottom) to work, so for most situations the tag-evaluation is the most efficient. The algorithm can be seen as two separate parts, the VCG Creation algorithm and the IICE profile generation algorithm.

4.2.1 **VCG creation Input**

To run the IICE algorithm’s first section, the VCG creation, the user must provide two data sets. The “subject” data set contains the individual or individuals who are going to be assigned a VCG. The “source” data set contains the data from which the VCG will be created. The data sets need to be in .csv format, with columns separated by a comma. The data sets have to have a common ID-column which will be identified as such by the user.

Once the data sets have been entered, they are compared for common column names, and a toggle-button is generated for each common column. If pressed, the column represented by the button will be included in the eventual VCG data.

The user then specifies the gender to extract from both data sets. If the user wants to run both genders, a second iteration of the algorithm can be initiated using the cell structure of Mathematica. As soon as the cell determining the gender has finished evaluating, the user can change the gender and reactivate the tag evaluation.

The user can also choose which matching variables to use for the k-nearest-neighbor matching, and how large the VCG should be. The default size is 100.

**Figure 17:** The VCG-X of a subject. The blue dot shows the patient values, the red dots show the current VCGs’ values and the white dots show the UKBB members of the same gender as the subject. The UKBB is expected to be viewed en masse, and so was given the less eye-catching color and size to allow the VCG and patient to stand out. The patient ID has been hidden.

\[\text{ppXX} = \text{propensity for XX}; \quad \text{CHD} = \text{Coronary Heart Disease}; \quad \text{T2D} = \text{Type 2 Diabetes}; \quad \text{srCOPD} = \text{self-reported COPD}; \quad \text{VCG} = \text{Virtual Control Group}; \quad \text{UKBB} = \text{UK Biobank}\]
For a faster procedure, the BCP plot creation can be inactivated by changing the “bcpcreate” parameter from 1 to 0. This procedure skips creating and saving the .pdf file BCP for each individual, which reduces the algorithm run time. The purpose of this option however is if the data set does not contain all BCP variables.

4.2.2 Initialization cell
The subject data set and the source data set are entered into a Mathematica initialization cell. This cell will prompt to be evaluated if the user attempts to evaluate any other cell unless the initialization cell has already been run in the current session. The initialization cell loads the two data sets, extracts the common columns, sets the name of the ID column, and sets the paths for the libraries and subprograms required. The cell also established the RLink, allowing R commands to be evaluated within Mathematica.

4.2.1 Output
For each individual in the subject data set three files are generated; the VCG data, the VCG reference data and the BCP plot. The user must, upon first use, edit the target directory for the algorithm to save the output. Based on the matching variables chosen, the name of the subject and VCG source data sets, sub-directories are generated within the main target directory, one for BCP plots, one for VCG data and one for the VCG reference data.

4.2.2 The VCG creation algorithm
Once all inputs and settings are set, the user can choose to evaluate all cells through the tag evaluation or evaluate the cells one by one.

The first cell locks the matching variables, the directories into which to save the output and the size of the VCG. If the subject data set is named “selfreported”, the VCG source data set is named “UKBB” and the main directory is:

C:/Users/user1/Mathematica/Mathematica-experiment/

Then the three sub-directories for matching variables ASATi and VATi will be:

C:/Users/user1/Mathematica/Mathematica-experiment/BCPdata/selfreported_UKBB/asati_vati
C:/Users/user1/Mathematica/Mathematica-experiment/visreferencedata/ selfreported_UKBB/asati_vati
C:/Users/user1/Mathematica/Mathematica-experiment/referencevalues/selfreported_UKBB/asati_vati

The second cell cleans the subject data to only contain subjects of the select gender.

The third cell clusters the members of the data sets based on the chosen matching variables using a KNN algorithm (3.4.3). For each subject, the 100 source data members are set to be their VCG (unless the user changed the VCG size). These 100 members are marked as “virtual control group” while the remaining source data set members are tagged with “population”.

The fourth cell calculates the median, the VCGs’ 75th and 25th quantiles and effect size of subject to VCG for each variable chosen from the common column toggle-button table. This, together with the subjects’ variable value constitutes the reference values. The reference values are saved as a RDS, one for each subject. The VCG data is also saved separately for each subject.

The fifth cell is the BCP creation, in which the BCP variables are set. The values from the VCG and patient are transformed using a logarithmic sigmoid transfer function to fit the BCP axes. The BCP data for all subjects is saved into a .csv file. For more detailed information about the BCP plot, see (21).

The sixth cell is a call to Python, and initiates the BCP plotting script, using the newly saved BCP data as input. The patient plots are saved as .pdf files individually into the BCP directory.

Once the sixth cell has finished evaluating, the VCG creation is complete, and the IICE profile can be generated.
4.2.3 Selecting the individual
IICE is built to handle availability of several data sets. The first cell in the IICE profile generation takes a directory as input and outputs all sub-directories available into a drop-down menu, cleaning the directory names from all parent directories. If the directory is

\[ C:/Users/user1/Mathematica/Mathematica-experiment/BCPData \]

and it contains the two sub-directories

\[ C:/Users/user1/Mathematica/Mathematica-experiment/BCPData/TypeA_UKBB \]

\[ C:/Users/user1/Mathematica/Mathematica-experiment/BCPData/TypeB_UKBB \]

then the drop-down menu will contain the options “TypeA_UKBB” and “TypeB_UKBB”. When one of these is selected the user can evaluate the second cell, which loads the original subject data set, opens the chosen data set directory (e.g Type A) and reads the subdirectories there. These subdirectories correspond to the different matching variables used in previously created VCGs. The process is the same as the data set selection in that the names are cleaned from the parent directories and a drop-down menu is presented to the user. If the chosen data set has been run for the matching variable combinations ASAT\textsuperscript{i}VAT\textsuperscript{i} and FR+IMAT+\textit{lff10p} then the drop down will contain the options “asati_vati” and “FR_imat_lff10p”.

Once the data set and matching variables have been chosen the user can evaluate the third cell. This cell will open the file for the first patient in the chosen directory and read the columns. These columns will be the same ones as the ones chosen initially when the VCG was created. A table of toggle-buttons will once again be generated, and the user can select which variables are of interest for the analysis being performed.

Finally, the user will choose which patient to study. Together with the button generation the third cell creates a drop-down menu of all patients present in the chosen data set subdirectory. The user can step through the different patients and is presented with eight chosen variable values that are extracted from the original subject data set that was loaded in the first cell. These variables are Gender, BMI, age, MR, VAT, ASAT, IMAT, lungfun, VLO\textsubscript{med} and ppsrCOPD.

When the user has chosen the individual to analyze, the fourth cell can be evaluated. This cell loads the data of the chosen individual created using the chosen matching variables. It also locks the variables chosen in the toggle-menu created from the previous cell. The VCG members are put in a separate data set using the tag “virtual control group” that was set during the VCG creation.

The fourth cell also contains a tag evaluation command, which evaluates all cells tagged with “IICE”. The IICE-cells are all the cells which contribute to the IICE profile of the chosen individual.

4.2.4 The IICE profile algorithm
There are 11 cells that contribute to the IICE profile.

The first cell creates the smooth density histogram plots for all chosen variables. It also calculates the effect size of the patient as compared to the reference population. The effect size of the subject to VCG is extracted from the reference values file, and the two values are added on the density histogram. The patient’s value is added by drawing a vertical line across the horizontal axis using the Epilog command. If the patient lacks the variable, the histogram is still plotted for the VCG and reference population, but with the text “The chosen patient does not have this information” across it. All density histogram plots are stored in a list named “histplots”.

The second cell generates the VCG-X. Since 3 plotf actors are necessary, these are by default set to ppCHD, ppT2D and ppsarcopenia. These three variables are generated through VCG analysis and as such are present for all
individuals, preventing start-up errors. The data corresponding to these variables are extracted for the individual, VCG and population and put into the 3D space. The cell contains settings for the plot style, which can be changed by the user if desired. Using the Control[] command the viewpoint can be changed to align with one of the variable axes, essentially transforming the VCG-X into a 2D scatter plot for the other two variables. The VCG-X is stored as the symbol “VCGX”.

The third cell generates the Patientsector. For each variable the patient value is divided by the median of the VCG and is entered into the SectorChart command. The Tooltip command is also used to add the previously created smooth density histogram of the variable as a mouse-over feature to the Patientsector. The Patientsector is stored as the symbol “patientsector”.

The fourth cell calculates the patient GOLD scale. If the patient has a lungfun value below 0.7 then their stage is calculated using the formula and ranking table described in 2.1.3. The intervals for the stages are calculated and added into separate lists. These lists are used as input for the NumberLinePlot[] command. The patient value is added as a vertical line using the Epilog command. The GOLD plot is stored as the symbol “GOLD.”

The fifth through ninth cells calculate the prevalence of five conditions in the current VCG and reference population. The conditions are sarcopenia, srCOPD, CHD, T2D and HCB. The occurrences are counted in both VCG and reference population, and an odds ratio is calculated. The percentage prevalences and the odds ratio are stored in the symbols “sarcopenia”, “selfreported”, “chd”, “t2d” and “hcb”.

The tenth cell uses the patient ID, VCG matching variables and chosen data set to load the correct BCP plot .pdf file into the symbol “BCPStar”.

The eleventh cell extracts the patient values for ppsrCOPD, and VLO_med and sets the cutoff values obtained through the AUROC analysis of ppsrCOPD and VLO_med. Depending on the gender of the patient the cutoff values are different. The Tooltip command is used for each VCG member to display their ID during a mouse-over.
4.3 VCG-based COPD diagnostics

In order to create a visualization for COPD risk using VCGs, the capability of assessing the risk using the available data had to be analyzed.

4.3.1 srCOPD propensity

Different matching variables for generating srCOPD propensity were tested through AUROC analysis. It quickly became apparent that the AUROC values were also affected by gender, and so the propensity algorithm was changed to handle different matching variables based on gender. The combination that resulted in the highest AUROC for women were VATi, wfr, IMAT, lungfun and ttvi. The highest AUROC for men was ASATi, MR, lungfun, lff10p, and ttvi.

Table 4: AUROC values for different matching variables for ppsrCOPD for both genders.

<table>
<thead>
<tr>
<th>Matching variables</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
<td></td>
</tr>
<tr>
<td>VATi, wfr, IMAT, lungfun, ttvi</td>
<td>0.7783</td>
</tr>
<tr>
<td>VATi, IMAT, lungfun, ttvi</td>
<td>0.7755</td>
</tr>
<tr>
<td>VATi, wfr, IMAT, lungfun, ttvi, MR</td>
<td>0.7671</td>
</tr>
<tr>
<td>VATi, wfr, ttvi, IMAT</td>
<td>0.73</td>
</tr>
<tr>
<td>VATi, lff10p, IMAT and ttvi</td>
<td>0.7296</td>
</tr>
<tr>
<td>VATi, lff10p and IMAT.</td>
<td>0.7057</td>
</tr>
<tr>
<td>ASATi, FR, MR</td>
<td>0.6746</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td></td>
</tr>
<tr>
<td>ASATi, MR, lungfun, lff10p, ttvi, wfr</td>
<td>0.7466</td>
</tr>
<tr>
<td>VATi, MR, lungfun, lff10p, ttvi</td>
<td>0.7413</td>
</tr>
<tr>
<td>VATi, MR, lungfun, lff10p, ttvi, FR</td>
<td>0.7378</td>
</tr>
<tr>
<td>ASATi, FR, MR, ttvi</td>
<td>0.6628</td>
</tr>
<tr>
<td>VATi, lff10p, IMAT and ttvi</td>
<td>0.6567</td>
</tr>
<tr>
<td>ASATi, FR, MR</td>
<td>0.6487</td>
</tr>
<tr>
<td>VATi, lff10p and IMAT.</td>
<td>0.6183</td>
</tr>
<tr>
<td>ASATi, MR, lungfun, lff10p, ttvi, wfr</td>
<td>0.7466</td>
</tr>
</tbody>
</table>
4.3.2 Correlation and ROC for VLO\textsubscript{med}, ppsrCOPD and GOLD\text{continuous}

Scatter plots and Pearson Correlation values for the continuous GOLD and VLO\textsubscript{med} against ppsrCOPD and HCB propensity can be seen in Figure 18. The ROC curves of VLO\textsubscript{med}, continuous GOLD and ppsrCOPD for diagnosing srCOPD can be seen in Figure 19.

Figure 18: Scatter plots for the two continuous FEV\textsubscript{1}-based lung assessments VLO\textsubscript{med} and GOLD\text{continuous} set against the propensity for srCOPD and the ppHCB. To the left are plots for the male UKBB data set population and to the right female.

FEV\textsubscript{1} = Forced Expiratory Volume first second; VLO\textsubscript{med} = Virtual Lung Obstruction median; GOLD\text{continuous} = transformation of the GOLD scale into a continuous spectrum; ppXX = propensity for XX; HCB = Health Care Burden, number of recorded hospital nights prior to scan; srCOPD = self-reported COPD; UKBB = UK Biobank.
Figure 19: ROC curves for srCOPD diagnosis using the continuous version of the traditional lung obstruction (GOLDcontinuous), virtual lung obstruction (VLO\text{med}) and propensity for srCOPD (ppsrCOPD) in the UKBB data set.

- \text{VLO}_{\text{med}} = \text{Virtual Lung Obstruction median; \text{GOLDcontinuous} = transformation of the GOLD scale into a continuous spektrum; \text{ppXX} = propensity for XX; \text{srCOPD} = self-reported COPD; \text{UKBB} = \text{UK Biobank.}
Finally, a sectorplot was developed that plots the patient and their VCG values for ppsrCOPD and VLO\textsubscript{med} (Figure 20). The cutoff values obtained through the ROC analysis are visualized as lines across the scatter plot space, separating the area into four, with the center marked with a cross where the cutoffs meet. Hovering the mouse over each member of the VCG displays their ID, in order to allow further analysis of that subject.

4.3.1 The COPD estimation

The COPD estimation is a scatter plot of the general COPD indicators ppsrCOPD and VLO\textsubscript{med}. The plotted populations are the individual and the VCG. The statistically derived threshold values are displayed as black lines across their respective axes. During mouseover of a VCG member, the ID of that member is displayed. When the cursor is placed in empty plot space the GOLD visualization is shown (Figure 21).
4.4 The IICE Tool

The IICE tool is a conceptualization and as such much of the code is still highly visible at the front end. The tool can be divided into two sections; VCG creation and IICE profile analysis.

4.4.1 VCG creation

In VCG creation the user chooses the data set of one or more individuals for which to generate a VCG, as well as choosing a data set from which to pull the VCG members. The user also chooses the matching variables based on which analysis to perform on the individual and the VCG, and whether or not the generate BCP stars. The user can then execute the remaining cells of the section and the algorithm will produce a VCG file and a BCP plot for each subject.

4.4.2 The IICE profile surface

In the second section of IICE, the user is presented with menus for selecting the desired data set, VCG variable matchings, which variables to plot and finally which subject to analyze. When the user evaluates the selection the visualization algorithm will initiate. The user will be presented with the chosen subjects’ IICE profile surface (Figure 22). The IICE profile consist of several visualization tools; the Patientsector, the COPD estimation, disease prevalences, the BCP star and the VCG-X.

Figure 21: The COPD estimation with GOLD mouseover functionality.

\( VLO_{\text{med}} \) = Virtual Lung Obstruction median; \( ppXX \) = propensity for XX; \( srCOPD \) = self-reported COPD; \( FEV_1 \) = Forced Expiratory Volume first second;
Figure 22: The IICE profile of an individual, using matching variables ASATi, lff10p, MR and VATi. Here the user can see the patient BCP, Patientsector, condition prevalence, COPD estimation and finally the VCG-X. The individuals’ ID has been hidden.

**ASATi** = Abdominal Subcutaneous Adipose Tissue; **VATi** = Visceral Adipose Tissue; **IMAT** = Intramuscular Adipose Tissue; **MR** = Muscle Ratio; **lff10p** = Liver Proton Density Fat Fraction (10 points); **FR** = Fat Ratio; **srCOPD** = self-reported COPD; **VLOmed** = Virtual Lung Obstruction median; **CHD** = Coronary Heart Disease; **T2D** = Type 2 Diabetes; **HCB** = Health Care Burden, number of hospital nights before scan.
4.4.3 The Patientsector

The most general of the visualization tools used is the Patientsector. It gives an overview of the patient values as compared to the current VCG median. This immediately provides the viewer an indication of which values seem abnormal, but not to which degree. To remedy this shortcoming, a mouseover over each of the variables will display that variable’s probability density plot, providing additional distribution information of that variable.

4.4.4 The BCP

The BCP is a standardized visualization of individual body composition. IICE imports the Python-generated image into Mathematica to display with the other visualizations. The BCP is described in detail in (21).

4.4.5 VCG-X

The VCG-X allows the user to explore and compare all available variables on several levels. Firstly, the distribution across variables for the VCG can indicate a close or distant relationship between the studied variable and the matching variables. Secondly, the VCG and individual can be compared to each other as well as to the reference population. Instead of single-variable analysis which is available through the Patientsector, the VCG-X allows the user to study the relationality of different variable values for the different individuals and populations.

Three drop-down menus allow the user to replace each of the three displayed variables. To reduce the strain of continuous updating, an “Update”-button is pressed to refresh the VCG-X once the variables have been changed.

Figure 23: The Patientsector with the mouse pointing at the asati variable. The mouseover effect is activated, displaying the asati probability density plot of that variable for the VCG and the same-gendered members of the UKBB. The patient value is indicated by the black vertical line.

ASATi; Abdominal Subcutaneous Adipose Tissue index; UKBB = UK Biobank; VCG = Virtual Control Group.
There is also a drop-down menu which allows the user to change between a set of viewpoints, including the option of studying the 3D scatter down one of the axes, essentially transforming the plot into a 2D scatter plot.

Figure 24: Three views of the VCG-X. The left, larger view is the default view set by Mathematica. The right examples display the same data, but from the viewpoint of one of the axes. The top right plot show the individual, VCG and reference population for the variables ppsarcopenia and ppT2D, while the lower plot show the variables ppCHD and ppT2D. The name of the third axis remains visible since the viewpoint does not edit the graphic beyond the viewed angle.

*ppXX* = propensity for XX; *CHD* = Coronary Heart Disease; *T2D* = Type 2 Diabetes;
4.5 Characterizing COPD

srCOPD members were scrutinized, first through BCP characteristics and then values for ppCHD, ppT2D, ppHCB, ppsarcopenia, VLOmed and lungfun.

4.5.1 At the BCP level

For the men, there were four BCP characteristics visually identified, appearing in varying prevalence. These were referred to as male Type A (23 subjects), Type B (11 subjects), Type C (3 subjects) and Type D (16 subjects). The characteristics are described and viewed in Figure 25.

Figure 25: The 4 male BCP types within the srCOPD data set. The examples shown are taken from individuals; each type can vary somewhat in appearance. Type A: A bloated appearance apart from the low IMAT. Elevated values of VATi, FR and lff10p in particular. Type B: Slightly bloated appearance, with lff10p values often lower than the other ectopic variables, as seen above. Type C: Normal or low values of all BCP variables except for a high IMAT. Type C is rare in the male SR data set. Type D: All values similar to the MDF.

ASATi = Abdominal Subcutaneous Adipose Tissue; VATi = Visceral Adipose Tissue; IMAT = Intramuscular Adipose Tissue; MR = Muscle Ratio; lff10p = Liver Proton Density Fat Fraction (10 points); FR = Fat Ratio.
For the women, four types were also identified. These were referred to as female Type A (14 subjects), Type B (10 subjects), Type C (16 subjects) and Type D (11 subjects). The characteristics are described and viewed in Figure 26.

Figure 26: The 4 female BCP types within the self-reported data set. The examples shown are taken from individuals; each type can vary somewhat in appearance. **Type A:** A bloated appearance for all BCP variables. **Type B:** The female Type B is in some instances similar to the male Type A, with low IMAT in relation to VATi and lff10p. VATi is however most often lower than in the male counterpart. **Type C:** Normal or low values of all BCP variables except IMAT. Type C is similar to its male counterpart, but more prominent in the female data set. Often the subjects have high values for FR, MR and ATATi. **Type D:** All values similar to the mdf reference.

**ASATi** = Abdominal Subcutaneous Adipose Tissue; **VATi** = Visceral Adipose Tissue; **IMAT** = Intramuscular Adipose Tissue; **MR** = Muscle Ratio; **lff10p** = Liver Proton Density Fat Fraction (10 points); **FR** = Fat Ratio.
4.5.2 Further variable correlation analysis

When analyzing the BCP characteristics some variables appeared more consistent than others. To further investigate, Pearson correlation analysis was performed for all BCP variables against ppsrCOPD. The variables with a correlation of more than R=0.6 for either gender is displayed in Figure 27.

4.5.3 At a comorbidity level

To explore differences in propensity for other conditions, 8 subsets of the UKBB data set were created. Four subsets consisted of the visually discernable BCP types in the self-reported group, i.e male and female type A, B, C and D. The other four subsets consisted of subjects with high ppsrCOPD and low values of another propensity value and can be seen in Table 5. The BCPs of each subset was scrutinized for similarities/differences, and then plotted in scatter plots for the other propensities, e.g the High ppsrCOPD-Low ppsarcopenia subgroup was analyzed in a T2D-, CHD- and HCB-propensity space. The purpose of this exercise was to separate dissimilar BCPs within the subgroups into smaller subgroups, based on potential comorbidity. Other variables were also screened. See APPENDIX B for the BCP’s of the groups.
Table 5: Subsets of subjects with high propensity for sCOPD. The groups were not cleaned from outliers when the min-max range was created.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Description</th>
<th>Size</th>
<th>ppsrCOPD values median, mean (min-max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type A</td>
<td>Described in 4.5.1</td>
<td>M: 23</td>
<td><strong>M: 0.0264, 0.0250 (0.0117-0.0359)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F: 14</td>
<td><strong>F: 0.0412, 0.0424 (0.0270-0.0536)</strong></td>
</tr>
<tr>
<td>Type B</td>
<td>Described in 4.5.1</td>
<td>M: 11</td>
<td><strong>M: 0.0249, 0.0302 (0.0161-0.0482)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F: 10</td>
<td><strong>F: 0.0220, 0.0222 (0.0101-0.0381)</strong></td>
</tr>
<tr>
<td>Type C</td>
<td>Described in 4.5.1</td>
<td>M: 3</td>
<td><strong>M: 0.0311, 0.0285 (0.0216-0.0329)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F: 16</td>
<td><strong>F: 0.0161, 0.0187 (0.0093-0.0392)</strong></td>
</tr>
<tr>
<td>Type D</td>
<td>Described in 4.5.1</td>
<td>M: 16</td>
<td><strong>M: 0.0199, 0.0202 (0.010-0.0405)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F: 11</td>
<td><strong>F: 0.0122, 0.0138 (0.010-0.0224)</strong></td>
</tr>
<tr>
<td>High ppsrCOPD</td>
<td>Low ppCHD</td>
<td>M: 29</td>
<td><strong>M: 0.0360, 0.0373 (0.0273-0.0519)</strong></td>
</tr>
<tr>
<td></td>
<td>Lowest 10 % of Coronary heart disease propensity</td>
<td>F: 32</td>
<td><strong>F: 0.0360, 0.0358 (0.0295-0.0418)</strong></td>
</tr>
<tr>
<td></td>
<td>out of highest 10% ppsrCOPD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High ppsrCOPD</td>
<td>Low ppT2D</td>
<td>M: 29</td>
<td><strong>M: 0.0375, 0.0367 (0.0278-0.0521)</strong></td>
</tr>
<tr>
<td></td>
<td>Lowest 10 % of Type 2 diabetes propensity</td>
<td>F: 32</td>
<td><strong>F: 0.0359, 0.0355 (0.0298-0.0418)</strong></td>
</tr>
<tr>
<td></td>
<td>out of highest 10% ppsrCOPD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High ppsrCOPD</td>
<td>Low ppsarcopenia</td>
<td>M: 29</td>
<td><strong>M: 0.0317, 0.0332 (0.0274-0.0474)</strong></td>
</tr>
<tr>
<td></td>
<td>Lowest 10 % of sarcopenia propensity out</td>
<td>F: 32</td>
<td><strong>F: 0.0378, 0.0381 (0.0295-0.0504)</strong></td>
</tr>
<tr>
<td></td>
<td>of highest 10% ppsrCOPD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High ppsrCOPD</td>
<td>Low ppHCB</td>
<td>M: 29</td>
<td><strong>M: 0.0360, 0.0353 (0.0273-0.0474)</strong></td>
</tr>
<tr>
<td></td>
<td>Lowest 10 % of HCB propensity out</td>
<td>F: 32</td>
<td><strong>F: 0.0358, 0.0358 (0.0294-0.0513)</strong></td>
</tr>
<tr>
<td></td>
<td>of highest 10% ppsrCOPD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 4.5.3.1 Type A, B, C and D

The self-reported cases were analyzed in 2D scatter plots that are available in APPENDIX A. The analyzed variables were lungfun, VLO\textsubscript{med} and propensities for CHD, T2D and sarcopenia. By analyzing the differences in mean of the UKBB data set and the data sets for all variables tested, Table 6 was constructed. A 95 % confidence interval was used to determine significance.

The four Types were also analyzed together in the VCG-X to more visually study their propensity distributions and respiratory condition as compared to each other (Figure 28).

Table 6: A table of 95 % CI mean difference tests for Type A-D as compared to the UKBB data set. If the difference places the data set on the less healthy side, it is marked with red. If the difference places the data set on the healthier side, it is marked with green. Non-significant differences are marked with a “0” with white background.

<table>
<thead>
<tr>
<th>Type (size)</th>
<th>ppCHD</th>
<th>ppT2D</th>
<th>ppHCB</th>
<th>ppsarcopenia</th>
<th>lungfun</th>
<th>VLO\textsubscript{med}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (23)</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B (11)</td>
<td>+</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>C (3)</td>
<td>0</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>D (16)</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (14)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B (10)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C (16)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>D (11)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 28: **Top**: VCG-X for Type A-D for both genders in the propensity spectrum for srCOPD, CHD and T2D. The scaling of the axis was set to include all subjects. The ppCHD for men is higher in scale than for the women. **Bottom**: Lung function and VLO\textsubscript{med} for Types A-D

ppXX = propensity for XX; CHD = Coronary Heart Disease; T2D = Type 2 Diabetes; srCOPD = self-reported COPD; VLO\textsubscript{med} = Virtual Lung Obstruction median; lungfun = Lung function (FEV\textsubscript{1}/FVC); FEV\textsubscript{1} = Forced Expiratory Volume first second; FVC = Forced Vital Capacity.
### 4.5.3.2 Propensity-based subsets

The four sub-sets created through the lowest 10% of propensities were tested for other propensities through a 95% CI. This can be seen in Table 7. The presence of Type A-D was also checked for in the low-propensity subsets, and the result can be seen in Table 8.

Table 7: Results from 95% mean difference confidence interval between the sub-sets for low ppT2D, ppCHD, ppHCB and ppsarcopenia. The sub-sets were tested against the UKBB data set. If the difference places the data set on the less healthy side, it is marked with red. If the difference places the data set on the healthier side, it is marked with green. Non-significant differences are marked with a “0” with white background.

<table>
<thead>
<tr>
<th>Type</th>
<th>ppCHD</th>
<th>ppT2D</th>
<th>ppHCB</th>
<th>ppsarcopenia</th>
<th>lungfun</th>
<th>VLO\text{med}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low ppCHD</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Low ppT2d</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Low ppHCB</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Low ppsarcopenia</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low ppCHD</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Low ppT2d</td>
<td>0</td>
<td>-</td>
<td>+</td>
<td>0</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Low ppHCB</td>
<td>0</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Low ppsarcopenia</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 8: Prevalence of Type A-D in the low-propensity sub-sets. The presence of the Types was determined visually. If a Type represents more than 25% it is marked with bright red, and more than 75 % in dark red. Zeros are marked with blue.

<table>
<thead>
<tr>
<th>Sub-set (subjects with BCP)</th>
<th>Type A</th>
<th>Type B</th>
<th>Type C</th>
<th>Type D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low ppCHD (27)</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>Low ppT2D (27)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Low ppHCB (26)</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Low ppsarcopenia (27)</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low ppCHD (32)</td>
<td>1</td>
<td>6</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Low ppT2d (32)</td>
<td>0</td>
<td>7</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>Low ppHCB (31)</td>
<td>15</td>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Low ppsarcopenia (31)</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

4.5.3.3 Low sarcopenia propensity

For the subjects with all values necessary to create the BCP-plot in the low-sarcopenia subgroup, the men were mainly of the Type A characteristic discovered in the BCP analysis of the srCOPD data set (20 out of 27). The remaining 7 men were similar to male type B (Slightly bloated for all BCP variables).

The women were most like the female type A characteristic discovered in the BCP analysis of the srCOPD data set (31 out of 31).

The high-ppsrCOPD, low-ppsarcopenia subset can be seen for each gender in a 3D scatter plot, for the propensities of T2D, CHD and HCB (Figure 29 and Figure 30)
Figure 29: The male subset for high ppsrCOPD and low ppsarcopenia plotted in a 3D scatter plot, together with all male members of the UKBB data set. Dimensions are propensities for CHD, HCB and T2D

\[ ppXX = \text{propensity for XX}; \ CHD = \text{Coronary Heart Disease}; \ T2D = \text{Type 2 Diabetes}; \ HCB = \text{Health Care Burden, hospital nights recorded before scan}; \ srCOPD = \text{self-reported COPD}. \]

Figure 30: The female subset for high ppsrCOPD and low ppsarcopenia plotted in a 3D scatter plot together with all female members of the UKBB data set. Dimensions are propensities for CHD, HCB and T2D.

\[ ppXX = \text{propensity for XX}; \ CHD = \text{Coronary Heart Disease}; \ T2D = \text{Type 2 Diabetes}; \ HCB = \text{Health Care Burden, number of recorded hospital nights prior to scan}; \ srCOPD = \text{self-reported COPD}. \]
Propensities for CHD and T2D were then plotted in a 2D scatter plot for facilitated viewing, together with their smooth density histograms (Figure 31 and Figure 32). Because of limited space, the subset was in the plot named “HpsrLpsar” for High ppsrCOPD and Low ppsarcopenia.

**Figure 31:** Scatter plot and smooth density histograms of female propensities for CHD and T2D in the high ppsrCOPD and low ppsarcopenia (HpsrLpsar) subset. Female UKBB data set for reference in white. Black line is drawn from point (0,0) to (1,1) to provide a linearity reference. Below each density histogram is the mean difference 95 % confidence interval for the variable comparing the displayed populations.

\[ ppXX = \text{propensity for XX; CHD = Coronary Heart Disease; T2D = Type 2 Diabetes; srCOPD = self-reported COPD; UKBB = UK Biobank} \]

**Figure 32:** Scatter plot and smooth density histograms of male propensities for CHD and T2D in the high ppsrCOPD and low ppsarcopenia subset. Male UKBB data set for reference in white. Black line is drawn from point (0,0) to (1,1) to provide a linearity reference. Below each density histogram is the mean difference 95 % confidence interval for the variable comparing the displayed populations.

\[ ppXX = \text{propensity for XX; CHD = Coronary Heart Disease; T2D = Type 2 Diabetes; srCOPD = self-reported COPD, UKBB = UK Biobank} \]
Figure 34: Clusters created when screening men with high ppsrCOPD and low ppsarcopenia using lungfun and lff10p. Male UKBB data set for reference in white. Below each density histogram is the mean difference 95 % confidence interval for the variable comparing the displayed populations.

\[ \text{Lff10p} = \text{Liver Proton Density Fat Fraction (10 points)}; \  \text{lungfun} = \text{Lung function (FEV1/FVC)}; \  \text{FEV1} = \text{Forced Expiratory Volume first second}; \  \text{FVC} = \text{Forced Vital Capacity}; \  \text{UKBB} = \text{UK Biobank}. \]

Figure 33: Clusters created when screening women with high ppsrCOPD and low ppsarcopenia using T2D propensity and lff10p, with only three outliers. Female UKBB data set for reference in white. Below each density histogram is the mean difference 95 % confidence interval for the variable comparing the displayed populations.

\[ \text{ppXX} = \text{propensity for XX}; \  \text{T2D} = \text{Type 2 Diabetes}; \  \text{srCOPD} = \text{self-reported COPD}; \  \text{lff10p} = \text{Liver Proton Density Fat Fraction (10 points)}; \  \text{UKBB} = \text{UK Biobank}. \]
When screening values other than propensities, lff10p and lungfun (FEV$_1$/FVC) were discovered to create clusters within the high ppsrCOPD-low ppsarcopenia for men (Figure 34).

The female subjects with High ppsrCOPD and low ppsarcopenia also separated using lff10p, which clusters when combined with propensity for T2D (Figure 33).

### 4.5.3.4 Low T2D propensity

In contrast to the low-sarcopenia-subgroup, nearly all subjects with all values necessary to create the BCP-plot in the low-T2D subgroup had small star-shaped BCPs similar to male type D (26 out of 27), with a slight tendency for relatively higher lff10p.

For the female subgroup, type B, C and D were present, but type B and C were most common (25 out of 32), with several subjects being similar to both.

The high-ppsrCOPD, low-ppT2D subset can be seen for each gender in the VCG-X, for the propensities of sarcopenia, CHD and HCB (Figure 35 and Figure 36).

Figure 35: The female subset for high ppsrCOPD and low ppT2D plotted in the VCG-X, together with all female members of the UKBB. Dimensions are propensities for CHD, HCB and sarcopenia

ppXX = propensity for XX; CHD = Coronary Heart Disease; HCB = Health Care Burden, number of recorded hospital nights prior to scan; srCOPD = self-reported COPD; T2D = Type 2 Diabetes; UKBB = UK Biobank.
Figure 36: The male subset for high ppsrCOPD and low ppT2D plotted in the VCG-X, together with all male members of the UKBB data set. Dimensions are propensities for CHD, HCB, and sarcopenia.

\[ \text{ppXX} = \text{propensity for XX}; \quad \text{CHD} = \text{Coronary Heart Disease}; \quad \text{HCB} = \text{Health Care Burden, number of recorded hospital nights prior to scan}; \quad \text{T2D} = \text{Type 2 Diabetes}; \quad \text{srCOPD} = \text{self-reported COPD}; \quad \text{UKBB} = \text{UK Biobank}. \]

Propensities for CHD and sarcopenia were then plotted in a 2D scatter plot for facilitated viewing, together with their smooth density histograms (Error! Reference source not found. and Figure 38). Due to limited space, the subset was in the plot named “HpsrLpt2d” for High ppsrCOPD and Low ppT2D.

Figure 37: Scatter plot and smooth density histograms of propensities for CHD and sarcopenia in female subjects with high ppsrCOPD and low ppT2D. Female UKBB data set for reference in white. Below each density histogram is the mean difference 95% confidence interval for the variable comparing the displayed populations.

\[ \text{ppXX} = \text{propensity for XX}; \quad \text{CHD} = \text{Coronary Heart Disease}; \quad \text{T2D} = \text{Type 2 Diabetes}; \quad \text{srCOPD} = \text{self-reported COPD}; \quad \text{UKBB} = \text{UK Biobank}. \]
Figure 38: Scatter plot and smooth density histograms of propensities for CHD and sarcopenia in male subjects with high ppsrCOPD and low ppT2D. Male UKBB data set for reference in white. Below each density histogram is the mean difference 95% confidence interval for the variable comparing the displayed populations.

ppXX = propensity for XX; CHD = Coronary Heart Disease; T2D = Type 2 Diabetes; srCOPD = self-reported COPD; UKBB = UK Biobank.

No clusters using other variables were distinguishable for the subjects with high ppsrCOPD and low ppT2D.

4.5.3.5 High ppsrCOPD, Low CHD propensity

The subgroup with high ppsrCOPD and low CHD propensity was found to contain several of the same subjects as the low T2D propensity subgroup (33 out of 59). The VCG-X for both genders displayed low propensity for T2D. 2D scatter plots display the propensities for HCB and sarcopenia (Figure 39 and Figure 40).
Figure 39: Scatter plot and smooth density histograms of female propensities for HCB and sarcopenia in subjects with high ppsrCOPD and low ppCHD. Female UKBB data set for reference in white. Below each density histogram is the mean difference 95% confidence interval for the variable comparing the displayed populations.

ppXX = propensity for XX; CHD = Coronary Heart Disease; HCB = Health Care Burden, number of recorded hospital nights prior to scan; srCOPD = self-reported COPD; UKBB = UK Biobank.

Figure 40: Scatter plot and smooth density histograms of male propensities for HCB and sarcopenia in subjects with high ppsrCOPD and low ppCHD. Male UKBB data set for reference in white. Below each density histogram is the mean difference 95% confidence interval for the variable comparing the displayed populations.

ppXX = propensity for XX; CHD = Coronary Heart Disease; HCB = Health Care Burden, number of recorded hospital nights prior to scan; srCOPD = self-reported COPD; UKBB = UK Biobank.
As with the similar group of high ppsrCOPD and low ppT2D, the low CHD group could not be visually clustered.

**4.5.3.6 High ppsrCOPD, Low HCB propensity**

The male low HCB propensity group BCPs were different from the types identified in the srCOPD data set, however some instances of male type A and D were observable in the male subset. For the women, all 4 types could be observed. The VCG-X for the CHD, T2D and sarcopenia propensities can be seen in Figure 41 and Figure 42.

---

**Note:**
- **ppXX** = propensity for XX; **CHD** = Coronary Heart Disease; **T2D** = Type 2 Diabetes; **srCOPD** = self-reported COPD; **HCB** = Health Care Burden, number of recorded hospital nights prior to scan; **UKBB** = UK Biobank.
Two clusters were found for women with high ppsrCOPD and low ppHCB when looking in a 2D scatter plot at sarcopenia propensity and CHD propensity (Figure 43). The clusters also appear in the UKBB data set, but in the opposite density distribution. The distribution of the men can be studied in Figure 44.

Figure 43: Women with high ppsrCOPD and low ppHCB when screening using CHD and sarcopenia propensity. The subjects form two distinct clusters, marked with circles. Female UKBB data set for reference in white. Below each density histogram is the mean difference 95 % confidence interval for the variable comparing the displayed populations

ppXX = propensity for XX; CHD = Coronary Heart Disease; srCOPD = self-reported COPD; HCB = Health Care Burden, number of recorded hospital nights prior to scan; UKBB = UK Biobank.

Figure 44: Scatter plot and smooth density histograms of male propensities for CHD and sarcopenia in subjects with high ppsrCOPD and low ppHCB. Male UKBB for reference in white. Below each density histogram is the mean difference 95 % confidence interval for the variable comparing the displayed populations

ppXX = propensity for XX; CHD = Coronary Heart Disease; srCOPD = self-reported COPD; HCB = Health Care Burden, number of recorded hospital nights prior to scan; UKBB = UK Biobank.

The men with high ppsrCOPD and low ppHCB could be clustered visually using ASATi and VATi (Figure 45).
4.6 INDIVIDUAL EXAMPLE ANALYSIS

To illustrate and evaluate the developed tools on an individual level, a subject was selected for individual analysis. The initial VCG for the analysis was made from the full UKBB data set using the matching variables ASATi, MR, VATi, IMAT and Iffl0p. These variables are all in some way related to the BCP variables without causing any variable overlap. The unused BCP variables are FR and ATATi. FR is calculated using both the fat compartment variables ASATi, VATi and the muscle variables rulb, lulb, rulf and lulf. All of these are already used either in themselves or through the variable MR. ATATi is the summation of VATi and ASATi, but since VATi is also in itself a BCP variable ASATi was used instead of ATATi.

The set of matching variables was chosen to explore which BCP variables would have the largest variety, to accentuate abnormal BCP-values of the subject. The reasoning for using the full UKBB data set was to maximize the number of available subjects to create as similar VCGs to the individuals as possible.

4.6.1 Individual – Female Type A

The example individual was selected from the Type A data set.

The selection steps are illustrated in Figure 46. In the final section of the data selection some of the individuals’ values are displayed. If the individual is missing one of these values this is also displayed here.

The patient can be seen to be female with an unusually high BMI. Her lung capacity seems normal, with a FEV1/FVC above 0.7 and a VLOmed around 1.

The patient values were put in the algorithm to generate her IICE profile. In the Patientsector there were three variables that stand out in particular: Iffl0p, ppsrCOPD and ppT2D (Figure 47). This was in spite of the fact that the VCG is created using all UKBB members, including subjects with high propensities for srCOPD and T2D.
Choose your data set and VCG source dataset
(dataset_VCGsourcedata):
TypeA_UKBB

Choose matching variables
(variable1_variable2):
asati_imat_lff10p_MR_vati

Select variables for Patientssector and VCG-X:

<table>
<thead>
<tr>
<th>age</th>
<th>height</th>
<th>weight</th>
<th>bmi</th>
<th>vat</th>
<th>asat</th>
<th>lff10p</th>
<th>FR</th>
<th>vati</th>
<th>asati</th>
<th>atati</th>
<th>imat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 1

age: 67.0027
BMI: 34.112
asati: 10.5034
imat: 15.023
VLOmed: 1.09434
psrsCOPD: 0.0407332

Figure 46: Data selection for analyzing the example individual from the data set Type A, matching on subjects from the UKBB data set using matching variables ASATi, IMAT, lff10p, MR and VATi

ASATi = Abdominal Subcutaneous Adipose Tissue; VATi = Visceral Adipose Tissue; IMAT = Intramuscular Adipose Tissue; MR = Muscle Ratio; lff10p = Liver Proton Density Fat Fraction (10 points); FR = Fat Ratio; ppXX = propensity for XX; CHD = Coronary Heart Disease; T2D = Type 2 Diabetes; TTVi = Total Thigh Volume index; srCOPD = self-reported COPD; VLOmed = Virtual Lung Obstruction median; UKBB = UK Biobank; FEV1 = Forced Expiratory Volume first second; FVC = Forced Vital Capacity.

tall variables, matching and non-matching, the individual was more similar to the VCG than the reference population. This conclusion is drawn by studying the value for “diff”, or the absolute difference between the individual’s effect sizes to the VCG and reference population. This includes spirometric variables which were not included in the matching.

The VATi, IMAT and ppCHD are all also above the VCG median, although not to the extent of lff10p, ppstrCOPD and ppT2D. The difference of VATi and IMAT is still of interest though since these, as well as lff10p, were used as matching variables.

The high level of ppstrCOPD can also be studied in the COPD estimation scatter plot. Here it can also be observed that not only is the individuals’ ppstrCOPD higher than her VCG median, she surpasses both threshold values in the COPD estimation plot, i.e her lungfun and VLOmed are both worse than the AUROC values for the two variables. The propensity for srCOPD appears high in the VCG, with all but one member above the threshold. The VLOmed is more evenly spread across the spectrum. The prevalence of disease in the VCG is also high, with Odds Ratios above 3 for both T2D, CHD and srCOPD. Since the VCG contains 100 subjects the percentages correspond to the amount of cases, so there are 5 people with self-reported COPD, 6 with CHD and 19 with T2D in the VCG.
Figure 47: The Patientsector of the example individual, displaying the density histograms of lff10p, ppsrCOPD, ppT2D, IMAT and VATi

VATi = Visceral Adipose Tissue; IMAT = Intramuscular Adipose Tissue; lff10p = Liver Proton Density Fat Fraction (10 points); ppXX = propensity for XX; T2D = Type 2 Diabetes; srCOPD = self-reported COPD; UKBB = UK Biobank.

Figure 48: The example individual in the COPD estimation with a UKBB-VCG matched using ASATi, IMAT, lff10p, MR and VATi

ppXX = propensity for XX; srCOPD = self-reported COPD; VLOmed = Virtual Lung Obstruction; CHD = Coronary Heart Disease; T2D = Type 2 Diabetes; HCB = Health Care Burden, number of recorded hospital nights prior to scan. UKBB = UK Biobank.
Looking at the patient BCP it can be observed that the VCG struggled to match most of the individual’s values to the point that she was placed outside or at the edge of the variables’ respective interquartile ranges. The exception is MR, where the VCG interquartile range envelop the individual (Figure 49).

Finally, the individual was analyzed in the VCG-X, using first the propensities for T2D, CHD and srCOPD and then the ectopic variables (Figure 51). These analyses make it apparent that the individual is an extreme case, located in the utmost edges of the subject spectrum for these dimensions.

As an additional reference the subject was given a healthy population VCG, containing subjects with no CHD, T2D, srCOPD or low lung function. The matching variables remained the same. The Patientsector had a similar yet reinforced appearance, with lff10p and ppT2D standing out in particular (Figure 50). Further analysis in the VCG-X of lff10p and ppT2D show that all VCG members with high ppT2D are also high in lff10p (Figure 50).

**Interpretation:** In health terms, the individual is at high propensity for T2D in particular, possibly due to her high levels of lff10p. Her propensity for CHD and srCOPD are also high. Her propensity for COPD has not yet resulted in reduced lung function, but the disease is common for her body type, with a 5% prevalence in a 100 subjects range.

---

**Figure 49:** BCP for the example individual, with a UKBB-VCG matched using ASATi, IMAT, lff10p, MR and VATi

ASATi = Abdominal Subcutaneous Adipose Tissue; VATi = Visceral Adipose Tissue; IMAT = Intramuscular Adipose Tissue; MR = Muscle Ratio; lff10p = Liver Proton Density Fat Fraction (10 points); FR = Fat Ratio; VCG = Virtual Control Group.
Figure 51: VCG-X screenshots for the example individual with her UKBB-VCG matched using ASATi, IMAT, lff10p, MR and VATi. The variables displayed are **Left**: ppT2D, ppCHD and ppsrCOPD and **Right**: IMAT, VATi and lff10p.

**ppXX** = propensity for XX; **CHD** = Coronary Heart Disease; **T2D** = Type 2 Diabetes; **srCOPD** = self-reported COPD; **VCG** = Virtual Control Group; **UKBB** = UK Biobank; **IMAT** = Intramuscular Adipose Tissue; **VATi** = Visceral Adipose Tissue; **lff10p** = Liver Proton Density Fat Fraction (10 points);

Figure 50: Patientsector and VCG-X for the example individual using a VCG from a healthy population. The Patientsector accentuates high lff10p values as well as a high propensity for T2D.

**ppXX** = propensity for XX; **T2D** = Type 2 Diabetes; **CHD** = Coronary Heart Disease; **VCG** = Virtual Control Group; **lff10p** = Liver Proton Density Fat Fraction (10 points); **srCOPD** = self-reported COPD; **extrememdf** = data set with subjects with no T2D, CHD presence, no FEV1/FVC below 0.7, and no srCOPD.
5 DISCUSSION

5.1 RESULTS

The results of this masters’ thesis consist of both visualization tools and their diagnostic performance. The diagnostic performance of VCG-based analysis is also discussed.

5.1.1 The tools

Each tool is evaluated by its technical properties and mainly their potential improvements. The tools are evaluated based on graphical integrity and excellence in a separate section in 5.3.

5.1.1.1 The Patientsector

The Patientsector compares the individual to the VCG, with the Tooltip option of also studying the VCG source reference population. It could prove beneficial to implement a MDF reference similar to the one displayed in the BCP star. The median circle could be replaced by the MDF while the median was displayed as an additional line along the individuals’ variable axis. This could compromise the readability of the visualization though.

Another alternative would be to plot an additional Patientsector in which the patient is compared to a MDF reference. As opposed to the BCP, this comparison could change based on the individual, by using a MDF VCG instead of set values.

As a third alternative, it would be interesting to return to the original design of the two-layered SectorChart and use the inner circle as the current VCG and the outer as either a comparison to a reference population or a healthy VCG created with the same matching variables as the inner circle. An additional drop-down menu could provide the user with the option of choosing both populations to compare with, or if only one population is of interest. This would further increase the dynamic properties of the Patientsector and allow for multiple VCG analysis in one plot. The comparisons between the two layers might be difficult due to scaling.

The placement of the variable names is automatic and are easily obscured if the patient value is much lower than the VCG median, or if the user is looking at many variables at once. It could increase readability if they were placed outside of the Patientsector, or in a locked position in relation to the median ring rather than the moving patient values. Another alternative would be to reintroduce the coloring used in the previous iterations of the plot and add a chart legend next to the Patientsector. This would make the Patientsector less grey-scale friendly, but the interactive nature of IICE already advocates digital use rather than print-out.

5.1.1.2 COPD estimation

The COPD estimation scatter plot could be developed further. The UKBB data set could be plotted to give an additional reference point, but it was considered unnecessary clutter since the cutoffs already served this purpose. Another alternative could be to replace the axes with a box plot of either the VCG or the UKBB data set of the same gender as the individual.

When hovering the mouse over the VCG members their ID is displayed. This could be further utilized by displaying additional individual information, such as their BCP star. Loading each members’ BCP star .pdf file would significantly increase the load time for the IICE generation.

The lung function value FEV₁/FVC should be considered to be used in the general COPD estimation. It could be used as a third dimension, i.e transforming the current 2D scatter plot consisting of ppsrCOPD and VLO_med into a 3D scatter plot similar to the VCG-X, but with added AUROC-cutoff planes. This could prove difficult to view, however, depending on the opacity required to distinguish the cutoff-planes. Another alternative is to allow the user to switch between these three variables and cutoffs in the current 2D environment.

Linköping University
SE-581 83 Linköping
013-28 10 00, www.liu.se
The cutoff values are already produced through data driven analysis, however they are dependent on the population and remain constant regardless of which individual is being analyzed. To increase the individualized nature of the COPD assessment, individual cutoff limits could be considered, possibly generated through analysis of several VCG’s.

An adjustment to facilitate reading of greyscale print-outs would be to replace the dot displaying the individual with a star or some other shape. This would further distinguish the patient from the VCG and reference population. The size difference is however considerable and should be distinct enough for such situations.

The GOLD stage plot which appears through Tooltip functionality for individuals whose FEV$_1$/FVC<0.7 uses a previously established COPD progression assessment. The visualization of the individual FEV$_1$ spectrum, however, adds previously hidden information. The width of the stages is displayed, and the user can make a more granulated assessment of pulmonary obstruction. The use of the vertical space could be used for something more intuitive than simply separating the stages.

The interesting find of high correlation between ppsrCOPD and both FR, VATi and MR in women should be further studied and could be used as to develop an additional COPD estimation variable.

5.1.1.3 VCG-X
The VCG-X has no Tooltip functionality. It is possible to add this functionality to the 3D space and as with the COPD estimation it could be interesting to be able to view the BCP’s of each VCG member. The required loading time discussed earlier could potentially increase further as these would be redistributed to each respective member each time the user changes a variable to study in the VCG-X.

The appearance of the VCG-X is different from the Patientsector and COPD estimation in color scheme. The blue background of the other visualizations was seen as more pleasant than the VCG-X grey. The VCG-X needed a background colors which allowed more colors to be distinguishable due to the three populations displayed; individual, VCG and reference population. A color analysis could be performed to create a more homogenous appearance when the VCG-X is combined with the other visualizations.

5.1.1.4 The BCP plotting
The BCP star is generated through a python script, which in this project was handled essentially as a black-box into which data was sent and a plot was generated into a .pdf file. Due to the complexity of the spoke dimensions in the BCP star and time constraints for this project it was decided to maintain the black-box approach.

To increase interactivity, such as introducing Tooltip or Mouseover functionality, the BCP star generation could be moved to Mathematica. This could also allow exploration of the changes in disease propensity of the subject by implementing Manipulate functionality on the BCP axes. For example, a Type C male is characterized mainly though their high IMAT value. By dragging the IMAT value to progressively smaller values and tracking the ppsrCOPD risk simultaneously could be an important visualization in a physician-patient interaction.

A user created package, RadarChart, allows for swift creation of such plots. Since the package is user created, future support is an uncertainty. An officially supported package published by Wolfram Research would be preferable.

5.1.2 VCG matching variables
The choice of with which variables to generate a VCG is a complex one and will differ based on the purpose of the VCG. There is the condition-specific approach, in which you would match solely on variables known to be connected to the condition of interest. For a subject with a high propensity to develop the condition, this would in theory match with the most condition cases. The second approach is to avoid the disease variables to study the patient's abnormalities as compared to the VCG. Another approach is the “wholesome” approach, in which you match on most or all metabolic variables available to generate a VCG as similar to the patient as possible. This allows
for a multi-condition analysis, as compared to the condition-specific approach. Often this method also pinpoints abnormal variables by the VCGs’ difficulty to match on it, which can be seen either through the spread of the VCG in the BCP or the Patientsector’s density histograms. The wholesome approach was used in the individual analysis examples (4.6).

It can also be argued that using calculated matching variables, such as ATATi, together with variables used in their calculation, such as VATi, will skew the result. The mentioned combination weighs the matching toward higher similarity in VATi value, since ATATi=VATi + ASATi. However, if VATi is of high importance to the analysis performed, it may be desired to weigh the matching towards that variable. Again, it is a matter of determining the purpose of the VCG being created.

5.1.3 The choice of matching data
The VCG can be created from any population with enough members and the same data columns as the individual. Depending on the purpose of the performed analysis it could in some cases be preferable to use a healthy population. The definition of a healthy population would in that case have to be addressed as well. In this project, the MDF data set required removal of members with low lung function and cases of self-reported COPD to be considered disease free. Other experiments might contribute further health requirements.

5.1.1 Virtual lung obstruction
Each individuals’ virtual lung obstruction level was calculated through a VCG for each UKBB data set subject using the matching variables FR, MR and ATATi. The reasoning for this was that these variables are less ectopic than the other three BCP variables yet have a strong metabolic relevance. Since it was later discovered that FR and MR appear to have a significant correlation to COPD, the virtual FEV1 may be lower than it could be, skewing the virtual lung obstruction value to be smaller for affected subjects.

The matching variables for this predictory variable should be optimized in a similar way to how the matching variables for ppsrCOPD was, possibly through a more automated approach. Still, the current ROC analysis shows equal or better results when using the VCG-predicted values than the traditional anthropometric formula.

5.1.2 The COPD BCP Types
The prevalence of Type A-D was not equal, and the homogeneity within each group was also varied. The groups had not been cleaned from outliers in propensity values when the min-max range was created. If this was done, the stratification in propensities for the Types could be improved.

5.1.2.1 The propensity-based sub-groups
The propensity-based sub-groups were internally similar in variables but could in some cases be clustered in other variables. This indicates that high-ppsrCOPD-subjects can be effectively stratified even further than into the four sub-groups created in this project. The low-CHD and low-T2D sub-groups were however so similar that they could be argued to be the same sub-group.

Most members in the high-ppsrCOPD, low-sarcopenia subgroup had an increased propensity for both T2D and CHD. This is likely due to the sarcopenia diagnosis of TTVi-thresholding. The subjects who receive a 1 in either DXA or MRI measurements are generally low in BCP values, and CHD and T2D are related to being overweight.

The ppsrCOPD intervals from minimum to maximum were almost identical for all groups, which suggests that the propensity for srCOPD is not particularly dependent on the other propensities analyzed.

It is interesting that the female subjects in the highest 10 % of ppsrCOPD and the lowest 10 % of ppT2D still are not significantly lower in ppT2D than the UKBB. This suggests a comorbidity between T2D and COPD.
It would be interesting to create further sub-groups, comprised of members with the highest propensities for sarcopenia, T2D, CHD and HCB.

5.1.2.2 Male Type A
The male Type A was highly prevalent in the self-reported data set and was easy to distinguish by the shape of the BCP. The range between the minimum and maximum ppsrCOPD values was smaller in male Type A than both Type B and D, even though the size of the group was about double. This is likely due to the high prevalence of srCOPD members of similar BCP characteristics.

When exploring other conditions in the high ppsrCOPD group, male Type A appeared almost exclusively in the high ppsrCOPD, low ppsarcopenia group. The ppsrCOPD median for the male Type A subset was the second highest of the four. There was however no decrease in lung function, or increase in virtual lung obstruction, which suggests that male Type A COPD patients generally maintain their lung function, or that their body composition changes as the disease progresses. This is further supported by a normal propensity for health care burden. These subjects might have COPD but they are generally able to maintain the disease without spending time at the hospital.

When comparing the Types (Figure 28) in the VCG-X Type A males are clustered in a comorbid area, with progressively increasing propensities for all three conditions.

In short: The male Type A COPD subject is generally not too burdened by their disease but has an increased propensity for T2D and possibly CHD. Their lung function is normal and their VLOmed is low.

In the previously described metabolic phenotypes there is no clear fit to male Type A.

5.1.2.3 Male Type B
The male Type B was less homogenic than Type A and could be argued to warrant further stratification. The range between the minimum and maximum ppsrCOPD values was the largest of the male groups, while the median was the second lowest after Type D.

The increased risk for CHD makes Type B unique among the men. The decrease in lung function could be explained either by progressed COPD in these patients or poor general health due to obesity as indicated by the elevated CHD propensity. One of the previously described phenotypes was characterized with sarcopenia, a redistribution of fat to the VAT compartment and reduced muscle mass.

In short: The male Type B COPD subject does not spend an unusual amount of time at the hospital but have an increased propensity for CHD and decreased lung function. It is possible that Type B has sarcopenia.

5.1.2.4 Male Type C
The Male Type C only consists of 3 subjects, but these 3 people have a high propensity for spending time at the hospital, their lungs are obstructed and their lung function is impaired.

The scarcity of male Type C cases could mean two things: either Type C is rare in COPD, or its’ members are prone to die. The high IMAT suggest reduced muscle mass and reduced muscle function, possibly due to cachexia. This description fits well with the Cachexic phenotype described previously.

In short: Male Type C COPD patients are possibly cachexic, with poor lung function caused by high levels of airway obstruction.

5.1.2.5 Male Type D
The male Type D express increased propensity for the TTVi-based sarcopenia diagnosis but are in general either as healthy or healthier than the other male UKBB data set members.
The male Type D has the lowest median ppsrCOPD value, but the largest range from minimum to maximum. The star shape is very common in the UKBB data set.

In the previously described metabolic phenotypes there is no clear fit to male Type D.

**In short:** Male Type D COPD patients appear healthy but are unusually prone to develop sarcopenia.

### 5.1.2.6 Female Type A

The ppsrCOPD range from minimum to maximum is the second smallest of the female types, and the median is almost double that of female Type B and almost four times the value of female Type D. Female Type A appears mostly in the low ppsarcopenia subgroups, which is expected due to their high propensities for both CHD, T2D and HCB.

Type A lung function and VLO$_{med}$ is normal and is possibly preserved due to their size. Their propensity and BCP variable values fit well to the previously described metabolic phenotype which was described to be obese with an increased cardiovascular risk and increase VAT and ASAT values.

**In short:** The female Type A have a high propensity for spending time at the hospital, and for developing CHD and T2D. Their general health appears to be poor, although their lung condition is normal.

### 5.1.2.7 Female Type B

The group has the second largest range between ppsrCOPD minimum and maximum, and the second highest ppsrCOPD median.

It is possible that this group would need further stratifications since the visual homogeneity of the group is low and no propensities separates it from the rest of the UKBB data set.

**In short:** The female Type B do not indicate increased propensities for CHD, T2D or HCB, and appear to have normal lung function with no severe obstruction. The group may require further stratification.

### 5.1.2.8 Female Type C

The group has the largest range between ppsrCOPD minimum and maximum, and the second lowest median. There are occurrences of typical female Type C in both the low ppCHD and ppT2D subgroups, but not in the low ppsarcopenia or low ppHCB.

While the group is similar in BCP characteristics to the male Type C, it does not share the male scarcity. It also lacks the same propensity for hospital visits and VLO$_{med}$ but retains the reduced lung function. Either the Type C characteristics are more common for the women, or it has a lower fatality rate. As with the similar male Type C, the female Type C is reminiscent of the previously described cachexic phenotype.

**In short:** The female Type C has no elevated propensities for CHD, T2D or HCB, but they do have a reduced lung function. As with the male Type C, these women might be cachexic.

### 5.1.2.9 Female Type D

The group has the shortest range between ppsrCOPD minimum and maximum, and the lowest median.

It is possible that female Type D manages to retain their lung functionality through a healthy lifestyle, based on the overall low BCP variable values and propensities.

**In short:** The female Type D appears to be extraordinarily healthy in terms of CHD, T2D and HCB propensities. Their lung function is normal, as is their propensity for sarcopenia.
5.1.3  The Algorithm

The algorithm contains a significant amount of RLink commands. In order to accelerate the algorithm, each VCG creation and IICE profile subsections should be analyzed for which language is most efficient in handling the task. In some cases matrices are generated in both languages which for larger data sets would cause reduced performance. Streamlining the data handling into one of the languages should be considered. However, by doing that the functionality and data interaction properties of one language would be lost. While Mathematica is more efficient in visualization, R seems to handle large data sets more efficiently, being capable of importing larger data sets at a significantly higher rate than Mathematica.

Interaction with the VCG algorithm is still in its early stages. All code is visible. To facilitate use of code that is rarely or never changed should not be as visible as it currently is, while aspects such as matching variables should be given a selection menu. The usage of tagged cells is already implemented, which enables the code to be hidden by minimizing tagged cells.

5.1.3.1  The input

Currently the user enters which data sets to use manually. If the tool was further developed for other users than experienced researchers then all data sets should be moved to a data set folder. The data set names would be read and translated into a drop-down selection menu similar to the patient selection. A similar solution could be used for the matching variable selection. The matching variable selection menu could be generated from the identified common columns.

5.1.3.2  The common columns

To compare columns in the subject data set and the VCG source data set the column names have to be identical. This may require the user to open the data sets externally to manually change the names of some columns that contain the same information as a column in the other data set but has e.g. capital letters or is delimited using dots instead of underscores. An algorithm could be developed to ignore capitalization and prompt the user to confirm or deny common column whose names are similar but not identical.

5.1.3.3  The output

The output could be optimized for size. A folder containing the VCG information for all VCG source data set members using the current method uses about 3 GB of data. For each subject, the selected data column for all same-gender members of the reference population is saved together with the VCG. This could be solved by during analysis matching the VCG to the VCG source data set and removing the matched rows, while keeping the rest as the reference population.

5.1.4  IICE profile

The IICE profile currently handles static data. It should be possible to add VCG-prediction functionality by allowing the user to adjust a patient's value for one of the matching variables and explore how disease prevalence and possibly propensity would change with that value. The VCG creation for one single individual only takes a few seconds at most, so generating new VCGs with adjusted values should be possible to do on the fly.

The VCG-X defaults to ppCHD, ppT2D and ppsarcopenia as variables to view. These should be changed to the top three variables in the chosen plot factors, since an error will occur if the user has excluded one of the default variables from the analysis.

5.1.5  The target demographic

The current tool requires knowledge about both WL, R, COPD and VCGs. The most appropriate user is a researcher with bio-technological background with some statistical knowledge.
5.2 Method
Most of the visualizations were developed prior to making any clinical analysis of the available data. As a result, the visualization tools have a wide variability but may be lacking in specificity for COPD analysis. The tools were however developed with the available data in mind, which means that functionality such as longitudinal analysis is not available, even though for example cachexia is generally diagnosed through a longitudinal depletion of muscle mass.

5.2.1 RStudio and Wolfram Mathematica
In this project it was attempted to perform all algorithmic actions through the Wolfram Mathematica framework. Most of the VCG creation algorithm is however performed through RLink using a custom R distribution. The option to use any other than the default R distribution is available only to the Windows operating system. If this compatibility issue remains unaddressed the VCG creation should to be implemented into pure R or translated completely to WL, while the analysis remains in Mathematica to retain the accessibility to interactive visualization. The IICE-profile RLink commands do not utilize the properties specific to the custom R distribution, and as such should not be affected. This would require the user to use both an R and Mathematica framework however, which could affect the efficiency of the system.

5.2.1 Diagnostic performance analysis of VCG
The diagnostic performance tests were mainly based in ROC analysis and Pearson correlation. The robustness of the tests could be analyzed through AUROC confidence intervals. The ROC plots were coded from scratch since no ROC analysis package was found within the Wolfram Language. The implementation of AUROC confidence interval calculation was not performed due to time constraints.

The translation of the GOLD scale into a continuous variable was necessary for these analyses, but this exercise could be considered altering the metric. More work could be done to establish a performance analysis which retains the discrete nature of the GOLD scale.

Traditional diagnosis techniques can be translated to be VCG-driven, which was shown through the creation of the VLOmed. The same approach could be used for the sarcopenia diagnosis of walk test performance and longitudinal muscle degradation if these data had been available.

5.2.2 Different plot options
Several additional plot types could be considered further, such as alternatives to color as a 4th dimension of multidimensional plots. The 3D scatter plot could for example have the point size adjusted in accordance with an additional variable, which could be seamlessly changed to another variable through a popup-menu. For this project the simplicity of fewer immediately visible dimensions were prioritized for facilitated surveyability. The complexity was instead incorporated through the Tooltip functionality; revealing more information about a specific area upon request instead of flooding the screen with data.

If the discussed Type A-D were further analyzed, Type-specific cutoff-analyses such as the COPD estimation scatter plot could be used to determine an individuals’ likeness to the different Types.

Some variables might both be considered metabolic and not, such as age. Should the VCG be matched towards your age group or is it more interesting to see if your VCG is older/younger than you, and introduce a “metabolic age”?

5.2.3 Rare phenotypes hidden in scarcity of data
The UKBB data set used in this thesis contains 6021 subjects. Out of these, 109 have self-reported bronchitis or emphysema. Depending on the number of actual phenotypes, 109 cases could weigh the diagnosis in favor of a more common phenotype. E.g, if there are only 10 cases of a rare phenotype that is divergent from a more common one, then the propensity value of those subjects will be low since the requirement in this project was for the propensity...
VCG to contain at least 20 cases. The VCG area would have to grow continually until 10 cases of another phenotype are incorporated, progressively lowering the propensity score.

5.2.4 The issue of self-reported COPD
Since the patients themselves have been allowed to enter whether or not they were diagnosed with emphysema or bronchitis, some may have entered it incorrectly. Others may not have entered a value at all, and subjects suffering from COPD may not have been diagnosed.

The size difference of the self-reported data set and the soft COPD data set made propensity for soft COPD to overdiagnose srCOPD, which might be preventing a useful diagnostic parameter. It seems unlikely that such a large group are undiagnosed however, and more likely that soft COPD is not a valid singular diagnosis tool.

5.3 Graphical consideration
One of the things to avoid during visualization is abbreviations. However, variable names such as “propensity for self-reported Chronic Obstructive Pulmonary Disease” did not fit in the plot windows and abbreviations were considered necessary.

There are several data visualization tools used in the IICE tool. In this section, each of them will be analyzed in the aspect of graphical excellence, and possible improvements will be discussed. The essence of graphical excellence is to illustrate complex ideas with clarity, precision and efficiency. The main considerations for each visualization for this exercise are:

➢ Is the data displayed complex enough to warrant the visualization?
➢ Does the visualization have a designated purpose?

5.3.1 The Patientsector
The Patientsector provides an overview of the patient as compared to the median of the current VCG. A well-matched VCG and a healthy subject would give a circular shape, which pleases the eye. But when that shape is broken, it is noticed immediately. This dynamic tool allows the user to view any number of variables, normalized to the same scale, and provides immediate information of which variables seem abnormal. Once that initial surge of information has been provided, the user can further scrutinize the variables that are deemed interesting, by hovering the mouse above them and seeing the density histogram plot of that variable for both the VCG and the same-gender population of the VCG source data set. The purpose of this visualization is to provide an introduction to the patient’s variable space, allowing for a more informed further analysis.

The Patientsector can be used for patients with a VCG created either from the complete UKBB same-gender population or any kind of subset, such as the MDF data set.

Further information could be added to the surface of the Patientsector, such as the maximum and minimum values of each variable, added through additional lines. Since these would not form a circular form, however, the graphical impact could hamper the expedious impressions provided by the Patientsector at first glance.

The Patientsector is capable of providing detailed information about any continuous variable present in the patient, VCG and a reference population, while using only a small circle of space. **The data complexity warrants the visualization.**

The purpose of the Patientsector is to give an immediate multi-variate appreciation of the patient as compared to the median of the current VCG, with seamless availability for deeper variable specific analysis.

An issue with the Patientsector is that goes against the principal of graphical integrity that states that the number of dimensions used to visualize data should be equal to the dimensions of the data. In the Patientsector, as a variable
increases in value, that variables’ sector grows outward. However, as the implied radius of the variable increases, so does the circumference. The documentation of whether the area is proportional to the variable growth is not documented in the WL documentation, but it has been assumed not to be. The solution could be to replace the “pie slices” for spokes, similarly to the BCP star. This might affect ease of comparison between variables though as the eye-pleasing circular form is removed.

5.3.2 COPD estimation
The sectorplot for ppsrCOPD and VLO_med uses statistically driven cut-off values to separate the space into 4, with low-low, high-low, low-high and high-high risk areas for the two COPD-related variables. This gives a general indication of both the health of the patient and VCG. The space used for this determination could be questioned as the patient’s relation to the cutoff values could be indicated by Boolean values. However, the addition of the VCG distribution as well as a visual indication of how remote the patient is from the thresholds adds valuable information to the user. For further information the VCG source population could be added to the plot as to give further reference, but the presence of the cutoff markers should already fulfill that purpose.

At first look the COPD estimation establishes a general risk for the individual. At further study the user can compare the individual to the VCG and determine if the patient appears abnormal in either variable. At the next level is the tooltip functionality. When a VCG member is below the cursor, their ID is shown. If the cursor is in empty plot space and the individual has a FEV1/FVC below 0.7, the GOLD scale specific to the studied individual is shown. In this way an additional spirometric dimension is introduced while retaining the comprehensibility of the 2D scatter plot.

Using a scatter plot to show the distribution one or more groups is an established visualization technique. The complexity is never expected to surpass a 2D space for population analysis, but the COPD estimation adds the threshold information, identification for each of the plotted members and an additional plot hidden in Tooltip functionality. The complexity of the data warrants the visualization method chosen.

The purpose of the COPD estimation is to provide a general disease risk assessment, facilitating further analysis of the patient using the other tools provided in IICE.

5.3.1 VCG-X
The VCG-X provides distribution analysis of three variables in multiple populations and serves as a VCG exploration tool. All variables are accessible through drop-down menus and can be easily interchanged by the user.

The VCG-X provides tri-dimensional analysis of the populations VCG and reference together with the individual. The space used is in 3D and the displayed variables can be replaced by others by the user, essentially limiting the analyzable dimensions only by the available data. The data complexity warrants the visualization used.

The purpose of the VCG-X is to provide accessibility of data exploration. The user chooses which variables to analyze and can potentially detect correlation of matching variables to non-matching variables by:

➢ The distribution of the VCG
➢ The extremity of both VCG and the individuals’ values as compared to a reference population
➢ The relationship of the individual and VCG.

5.3.2 The BCP
The BCP is a standardized assessment of the distribution of six MRI-measurable biomarkers, displaying both the individual, the VCG and a healthy reference population. It accomplishes presentation in a 2D plane of one or more populations, as well as the individual’s values, while putting them in the context of what values are healthy.
The standardization limits which variables to study but enables the longitudinal improvement of readability. While more dynamic tools such as the Patientsector would change in appearance each time the user changes which variables to analyze, experienced BCP analysts are able to assess health risks using only the BCP.

The complexity of the data expressed through the BCP includes not only the reference of population and MDF as compared to the individual, but also teases the eye if the star shape is broken, bringing attention to variables that stand out. Differences in ATATi, MR and FR are however more difficult to distinguish than the ectopic VATi, IMAT and lff10p since they even in a healthy person appear as edges, which can be seen as problematic. Still, the data complexity warrants the visualization used.

The purpose of the BCP is similar to the Patientsector, but the approach is more standardized. Mastering the BCP analysis enables the user to draw immediate conclusions of health risks.

5.4 Future work
In addition to the improvement strategies suggested in the above discussion topics, some future areas of research are presented below.

5.4.1 KOLD study validation
Due to the missing lff10p data in the COPD vitamin D study the data set was not used in the assessment of IICEs’ diagnostic performance. When the COPD study is complemented its’ data could be used to explore the prevalence of the identified Types in confirmed and progressed cases of COPD.

5.4.2 Definition of sarcopenia
The sarcopenia diagnosis method used in this project is based on a TTVi threshold. This threshold favors subjects of low overall volume, and not necessarily sarcopenic ones. A more robust diagnosis method of this condition would be valuable to the stratification of COPD using metabolic measurements.
6 CONCLUSION

The diagnostic performance of VCG based analysis is promising, outperforming the GOLD continuous variable using both ppsrCOPD and VLOmed. The effective clustering of high-ppsrCOPD subjects using both other propensities and metabolic variables suggest potential for a high-resolution, metabolism-based stratification of the COPD condition.

Based on these discoveries, the IICE tool provides an intuitive environment in which to analyze a single individual’s properties; metabolic, spirometric and other. The VCG provides information about how people who are similar to you in some respects look in others, as well as how both the VCG and you look compared to the rest of a larger population.

This master thesis suggests that VCG-based COPD analysis is equal or better than more traditional diagnostic techniques. Further optimization and higher data density for propensity values should further the diagnostic performance of the VCG-based techniques. It is also shown that a user interface taking up less space than a single standard screen allows for a faceted, detailed and multivariate analysis of individual COPD risk.

This tool allows for a high resolution metabolic analysis of COPD cases, enabling more individually suited treatment plans.
REFERENCES


23. UK Biobank [Internet]. [cited 2018 Jan 18]. Available from: http://www.ukbiobank.ac.uk/
APPENDIX A: SCATTER PLOTS OF TYPE A-D

Type A Men
Type A Women

- Scatter of gender F for:
  - Population 1: unbestended
  - Population 2: TypeA

- Scatter of gender F for:
  - Population 1: unbestended
  - Population 2: TypeA

- Scatter of gender F for:
  - Population 1: unbestended
  - Population 2: TypeA

- Scatter of gender F for:
  - Population 1: unbestended
  - Population 2: TypeA
Type B Men

Scatter of gender M for:
Population 1: untoextendd
Population 2: TypeB

prop_t2d_bcp

Scatter of gender M for:
Population 1: untoextendd
Population 2: TypeB

pppsarcopenia

Scatter of gender M for:
Population 1: untoextendd
Population 2: TypeB

VLOned
Type B Women

- prop_t2d_bcp
- ppsarcopenia
- VLOmed
Type C Women

Scatter of gender F for:
Population 1: unk_extended
Population 2: TypeC

Scatter of gender F for:
Population 1: unk_extended
Population 2: TypeC

Scatter of gender F for:
Population 1: unk_extended
Population 2: TypeC
Type D

Scatter of gender M for:
Population 1: ubbextended
Population 2: TypeD

Scatter of gender F for:
Population 1: ubbextended
Population 2: TypeD
Scatter of gender M for:
Population 1: ukbextended
Population 2: TypeD

Scatter of gender F for:
Population 1: ukbextended
Population 2: TypeD
APPENDIX B: BCPs FROM DATA SETS

Type A from Self-Reported set

Males

Females
Type B from Self-Reported set

Males

Females
Type C from Self-Reported set

Males

Females

Type D from Self-Reported set
Males

Females
High ppsrCOPD, Low sarcopenic propensity

Lowest 10% in sarcopenic propensity out of highest 10% ppsrCOPD: Females

Lowest 10% in sarcopenic propensity out of highest 10% ppsrCOPD: Males
High ppsrCOPD, Low T2D propensity

Lowest 10% in T2D propensity out of highest 10% ppsrCOPD: Females

Lowest 10% in T2D propensity out of highest 10% ppsrCOPD: Males
High ppsrCOPD, Low CHD propensity

Lowest 10% in CHD propensity out of highest 10% ppsrCOPD: Females

Lowest 10% in CHD propensity out of highest 10% ppsrCOPD: Males
High \( \text{ppsrCOPD}, \) Low HCB propensity

Lowest 10% in HCB propensity out of highest 10% ppsrCOPD: Females

Lowest 10% in HCB propensity out of highest 10% ppsrCOPD: Males