Nuno Filipe Ramalho da Cunha Tavares

Licenciado em Engenharia e Gestão Industrial

Multivariate Analysis Applied to Clinical Analysis Data

Dissertação para obtenção do Grau de Mestre em
Engenharia e Gestão Industrial

Orientador: Doutora Ana Sofia Leonardo Vilela de Matos,
Professora Auxiliar, FCT, UNL

Co-orientadores: Doutor Rogério Salema Araújo Puga Leal,
Professor Auxiliar, FCT, UNL

Júri:

Presidente: Professora Doutora Virgínia Helena Arimateia de Campos Machado

Vogais: Professor Doutor Duarte Paulo Martins Torres

Professora Doutora Ana Sofia Leonardo Vilela de Matos

Dra. Carla Alexandra Fino Alberto da Mota

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Março de 2014

Multivariate Analysis Applied to Clinical Analysis Data

Acknowledgements

The conclusion of this dissertation would have been impossible without the help from the people around me, to whom I would like to thank.

Firstly, my supervisors Ana Sofia Matos and Rogério Puga Leal, assistant professors from the Department of Mechanical and Industrial Engineering in *Faculdade de Ciencias e Tecnologia* of the *Universidade Nova de Lisboa*, for the commitment, support and guidance.

Also my coordinator at the National Health Institute, Carla Mota, from the Department of Nutrition of the National Health Institute Doutror Ricardo Jorge, for the knowledge transmitted and support in a field that was not my own.

To Isabel Castanheira, researcher from the Department of Nutrition of the National Health Institute Doutror Ricardo Jorge and the whole Department of Nutrition, for the excellent research and work atmosphere. And the National Health Institute Doutror Ricardo Jorge for the opportunity to perform this research.

Thank you to my university colleagues, for the friendship and support during my academic journey.

Last but definitely not least to my family, for the unconditional support and motivation, and particularly my parents for giving me the opportunity to study.

Abstract

Folate, vitamin B12, iron and hemoglobin are essential for metabolic functions in the body. The deficiency of these can be the cause of several known pathologies and, untreated, can be responsible for severe morbidity and even death. The objective of this study is to characterize a population, residing in the metropolitan area of Lisbon and Setubal, concerning serum levels of folate, vitamin B12, iron and hemoglobin, as well as finding evidence of correlations between these parameters and illnesses, mainly cardiovascular, gastrointestinal, neurological and anemia.

Clinical analysis data was collected and submitted to multivariate analysis. First the data was screened with Spearman correlation and Kruskal-Wallis analysis of variance to study correlations and variability between groups. To characterize the population, we used cluster analysis with Ward's linkage method. Finally a sensitivity analysis was performed to strengthen the results.

A positive correlation between iron with, ferritin and transferrin, and with hemoglobin was observed with the Spearman correlation. Kruskal-Wallis analysis of variance test showed significant differences between these biomarkers in persons aged 0 to 29, 30 to 59 and over 60 years old. Cluster analysis proved to be a useful tool when characterizing a population based on its biomarkers, showing evidence of low folate levels for the population in general, and hemoglobin levels below the reference values. Iron and vitamin B12 were within the reference range for most of the population. Low levels of the parameters were registered mainly in patients with cardiovascular, gastrointestinal, and neurological diseases and anemia.

Keywords: Cluster analysis, Spearman correlation, folate, vitamin B12, biomarkers, Kruskal-Wallis

Resumo

Ácido fólico, vitamina B12, ferro e hemoglobina são essenciais para as funções metabólicas no organismo. A falta destes pode ser causa de várias patologias conhecidas e, sem tratamento, podem levar a morbidade e até morte. O objetivo deste estudo é caracterizar uma população, residente na área metropolitana de Lisboa, em relação aos níveis séricos de ácido fólico, vitamina B12, ferro e hemoglobina, assim como encontrar indícios de correlações entre esses parâmetros e as situações clínicas autoreportadas.

Os dados, obtidos por análises clinicas, foram submetidos a uma análise multivariada. Posteriormente foram filtrados através de correlações de Spearman e análise de variância Kruskal -Wallis para estudar as correlações e existência de diferenças entre os grupos etários. Para caracterizar a população foi utilizada a análise de *clusters* com o método de ligação de Ward. Foi feita uma análise de sensibilidade para reforçar os resultados.

Através da correlação de Spearman foram encontradas correlações do ferro com a ferritina e transferrina e com a hemoglobina. A análise de variância Kruskla-Wallis evidencia a existência de diferenças significativas entre os biomarcadores e pessoas com idades entre os 0 e os 29 anos, entre os 30 e os 59 anos e com idade superior a 60 anos. A análise de *clusters* mostrou-se uma ferramenta útil para caracterização da população com base nos seus biomarcadores, revelando níveis em geral baixos de folatos na população e muitos valores de hemoglobina abaixo do valor de referência. O ferro e a vitamina B12 apresentam valores dentro dos limites na maior parte dos casos. Os níveis baixos nos parâmetros foram registados maioritariamente em pacientes com doença cardiovascular, gastrointestinal, neurológica e anemia.

Palavras-chave: Análise de *clusters*, correlação de Spearman, folatos, vitamina B12, biomarcadores, Kruskal-Wallis

Contents

Figures Index

Tables Index

List of Symbols

d – Distance

F_0 – Fisher statistic

H – Kruskal-Wallis statistic test

H_0 – Null hypothesis

H_1 – Alternative hypothesis

k – Number of groups

MS_B – Between mean squares

MS_W – Within mean squares

N – Total number of individuals

n – Sample size

SS_B – Between sum of squares

SS_E – Error sum of squares

SS_T – Total sum of squares

SS_W – Within sum of squares

$\bar{X}$ – Mean

α – Significance Level

ϑ – Degrees of freedom

ρ_S – Spearman correlation

σ – Standard deviation

χ^2 – Chi-square statistic

List of Abbreviations

ANOVA – Analysis of variances

d.f. – Degrees of freedom

DNA – Deoxyribonucleic acid

EURRECA – European micronutrients recommendations aligned

HCA – Hierarchical cluster analysis

MCW – Mean corpuscular volume

RNA – Ribonucleic acid

SPSS – Statistical product and service solutions

1. Introduction

1.1. Motivation

Iron, vitamin B12 and folate are essential for metabolic functions. The deficiency or lack of these nutrients can be the cause of several known pathologies and, untreated, can be responsible for severe morbidity and even death (Roddie & Davis 2009).

It is important to assess the state of these parameters in the Portuguese population to take measures, if necessary, to promote health programs, preventing diseases and improving the overall health state of the population.

The European Micronutrient Recommendations Aligned (EURRECA) project studied the food intake adequacy in European countries, for a range of status biomarkers, including folate, vitamin B12 and iron. In Portugal the data was obtained through food frequency questionnaires retrieved between 1999 and 2003. In that study iron and cobalamin intakes in the Portuguese population are adequate, with deficiency values under 7%, with higher percentages for the elderly population. Folate intakes however, prove to be inadequate for over 20% of the population, also with higher levels in elderly citizens. This fact led us to use folate as the factor for decision making in the study. The EURRECA study was conducted using a population from Porto, so it is not representative of our population, but still provides a picture of the Portuguese population (Viñas et al. 2011).

The analysis of the blood levels of the biomarkers can be submitted to statistical multivariate analysis techniques to characterize our study population.

In an increasingly competitive world, statistics play a decisive part in both industrial and scientific research. It is key for an engineer to know and master these techniques, being able to apply them to any problem in his path (Vining & Kowalski 2010).

1.2. Objectives

The objective of this exploratory, descriptive and transversal study is the characterization of the Portuguese population, with focus on the metropolitan area of Lisbon and Setubal, concerning serum levels of folate, vitamin B12, iron and hemoglobin. Clinical analysis data collected was collected from several counties in the Lisbon and Setubal districts in association with a Portuguese clinical analysis laboratory and National Health Institute Doutor Ricardo Jorge (INSA). The data will be submitted to multivariate analysis. We will try to find a valid correlation with information about age and health complications, and obtain an overall description of the population based on levels of folate, vitamin B12, iron and hemoglobin.

Folate will be the main parameter in study, and we will try assess the condition of the Portuguese population's serum folate concentrations.

This study will serve as grounding for future studies, mainly about folate supplementation and food fortification.

1.3. Document Structure

The dissertation is divided into five chapters, each referring to a different step in the study. At the end it's possible to consult the appendix files, containing tables and figures that help to assess some parts of the document.

This first chapter, consists of an introduction to the motivation and objectives for the study, and to the structure of the dissertation and its contents.

The second chapter is the literature review. It is an introduction to the biomarkers that are under study, their effects on the body, what illnesses they may cause and what factors affect their behavior. Also the methods used in the multivariate analysis, an introduction to the statistic tests, their methodology and their purpose.

The procedures and the methodology used are described in the third chapter. Here the methods are in chronological order, so the reader can better understand which method should be applied at what point. The procedures are also explained, as well as the reasons why they were used.

The fourth chapter contains both the results from the multivariate analysis and their respective discussion. In this section the results are presented and commented to help the reader interpret the outputs of the statistical tests.

Finally the fifth chapter contains the study's conclusions and also recommendations for future research.

2. Literature Review

2.1. Biomarkers

2.1.1. Folate

The term "folate" refers to the complete group of folic acid derivatives, including the polylglutamates naturally present in foods and folic acid that is a synthetic form used for food fortification and nutritional supplements. Folates are absorbed in the small intestine. Higher doses of folic acid are absorbed through a nonsaturable passive diffusion process (Iyer & Tomar 2009).

Folate is a B-group vitamin that participates in many metabolic pathways such as DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) biosynthesis and amino acid interconversions, being involved in essential functions of cell metabolism. Since mammalian cells cannot synthesize folate, an exogenous supply of this vitamin is necessary to prevent nutritional deficiency. Folate deficiencies occur frequently, even in well-developed countries, although folate is part of a normal human diet (Iyer & Tomar 2009).

Folate deficiency in the organism can be caused by malnutrition due to the intake of insufficient doses of foods containing folate but a more important risk factor is malabsorption. This happens especially in patients suffering from gastrointestinal complications such as Crohn's disease, tropical sprue and gluten sensitive enteropathy that cause a deficient absorption of the vitamin. Also some drugs such as oral contraceptives, anticonvulsants, H-2 receptor antagonists, barbiturates, cholestyramine, anti-inflammatory drugs, methotrexate, aspirin antacids and

alcohol can cause folate depletion. Folate deficiency may also be due to vitamin B12 deficiency (Iyer & Tomar 2009).

Folate deficiency can be the cause for several health problems, such as anemia, cancer, cardiovascular diseases as well as neural tube defects in newborns (Iyer & Tomar 2009).

2.1.2. Vitamin B12

Vitamin B12 or cobalamin exists only in foods of animal origin including milk, cheese or eggs and it resists cooking. It is synthesized by microorganisms and it's stored mainly in the liver(Roddie & Davis 2009). B12 acts as a coenzyme in the methylation of homocysteine to methionine, essential to the generation of metabolically active forms of folate that are required for DNA synthesis (Roddie & Davis 2009).

The causes for vitamin B12 deficiency may be dietary deficiency, ineffective absorption or metabolic inhibition. Strict vegans are at risk of levels low enough to cause megaloblastic anemia. Impaired absorption is commonly caused by pernicious anemia, gastrectomy, terminal ileal disease, bacterial overgrowth in the small intestine, alcoholism, fish tapeworm and tropical sprue. Deranged metabolism can also be a cause for low cobalamin (Roddie & Davis 2009).

Vitamin B12 deficiency is associated with high levels of homocysteine, an indicator of risk for heart disease, connecting B12 deficiency with anemia and arteriosclerosis. It's also a risk factor for coronary heart disease and stroke. Cobalamin deficiency is also associated with neurological illness (Christensen et al., 1999; Chui et al., 2001; Mccully, 1999).

2.1.3. Iron

Body iron is regulated by a complex system of proteins. Commonly the absorption of iron through food matches the daily losses from the gut skin and genitourinary tract. Menstruation, pregnancy, early infancy and adolescence increase the daily iron requirements. The absorption of iron can be increased, if necessary, to cope with such demands, but unless the intake and absorption keep up with the extra loss, deficiency will occur. On the other hand, there is no physiological mechanism for excreting iron surplus, iron accumulates in pathological states of increased gut absorption or following repeated blood transfusions (Roddie & Davis 2009).

Iron deficiency can be caused by an iron-poor diet, malabsorption, increased physiological demands or chronic blood loss, particularly in neoplastic gastrointestinal diseases. The commonest cause of anemia worldwide is iron deficiency anemia (Roddie & Davis 2009).

Excess of iron in the body can also be toxic, forming high oxidative free radicals in the presence of non-transferrin bound iron. It affects the liver, skin, endocrine glands and the heart (Roddie & Davis 2009).

2.1.4. Ferritin

Ferritin concentration is very low in the blood, being sometimes a poor indicator for body iron levels. This low concentrations decline particularly in iron deficiency, being detectable before any other manifestation or symptom, making it an effective status biomarker (Burtis et al. 2012).

2.1.5. Transferrin

Transferrin is the protein responsible for serum iron transport. High concentrations of transferrin indicate a deficiency in iron, possible cause for anemia, liver disease or malnutrition. This condition can also be observed during pregnancy and estrogen administration (Burtis et al. 2012).

2.1.6. Hemoglobin

Hemoglobin is a protein responsible for transporting oxygen in the blood. The two main types of hemoglobin complications are thalassemia, a deficiency in hemoglobin production and hemoglobinopathy, a deficiency in hemoglobin. Both conditions are cause for anemia (Burtis et al. 2012). Hemoglobin concentrations tend to lower with age, due to the bone marrow's inability to recover from injury (Lima-Costa 2012).

2.2. Population and Clinical Conditions

2.2.1. Demographics

The aging process of the European population started in the twentieth century, following the socioeconomic development and progress in medicine and health care systems. This phenomenon becomes more evident in developed countries (Mesquita 2009; Serviço de Estudos sobre a População Departamento de Estatísticas Censitárias e da População 2002).

The Portuguese population is growing older every year. The low birth rate, allied to a growing migratory tendency over recent years contribute to this increase in the elderly population (Instituto Nacional de Estatística 2011; Carrilho & Patrício 2004).

2.2.2. Elderly

Vitamin B12 deficiency is a common problem in the elderly population, although it's seldom diagnosed due to scarcity of evident symptoms (Andrès et al. 2007; Baik 1999).

Malabsorption from food, is the main cause for malabsorption, since the elderly patients' digestive system lacks the ability to synthesize the vitamin, although it can absorb the vitamin in the crystalline form. Another cause is pernicious anemia (Andrès et al. 2007; Baik 1999; Campbell et al. 2003).

Malabsorption can also be a consequence of gastric disorders or diseases like pancreatitis, Crohn's disease or gastric bypass surgery (Andrès et al., 2007).

Other studies suggest external factors like socio-economic gradient and education directly affect vitamin intake and diet quality. In these studies, persons with higher income or higher education tend to have a more nutritious diet. Also the quality of the diet is better for the elderly who share their meals with others instead of eating alone (Weimer 1998; Cabrera et al. 2007).

Low levels of hemoglobin due to lack of cobalamin and folate, are related to muscular and mental debilitation, resulting in decreased physical activity and cognitive difficulties. Elderly persons who suffer from folate and vitamin B12 deficiency seek medical attention more often than healthier ones (Lima-Costa 2012; Matthews 1995).

2.2.3. Mental Illness

The deficiency of folic-acid and vitamin B12 is a common condition in the elderly population. This deficiency contributed to high levels of serum homocysteine that is associated with cognitive deterioration and other mental conditions (Lancellotti 2012; Ramos et al. 2005; Mazeh et al. 2005).

Some studies show an inverse correlation between low levels of folate and vitamin B12 and Alzheimer's disease (Reynolds 2006; Clarke et al. 1998).

2.2.4. Pregnancy

Vitamin B12 and folate are important for the healthy fetus and prenatal development, affecting all areas of development from brain to bone. A deficiency of these nutrients can be the cause of neural tube defects in the antenatal period. Because of these factors, women are often subject to supplementation with both folate and vitamin B12 amongst others (Pepper & Black 2011; Czeizel & Vargha 2004).

This supplementation factor forced us to remove the pregnant women data from the current study.

2.2.5. Gastric and Liver Complications

Gastrointestinal diseases are often a cause for vitamin and nutrient malabsorption (Burtis et al. 2012). Patients submitted to gastric bypass surgery have folate, vitamin B12 and iron absorption problems, hence it's common for them to have low serum levels of these nutrients. There is no evidence that oral supplementation is a solution for this problem (Brolin et al. 1998).

Liver disease can affect nutrient intake and disposition (Burtis et al. 2012).

2.2.6. Smoking and Alcohol

Smoking and heavy drinking (alcoholism) have been related to low concentrations of serum folate, B12 and red blood cells (Okumura & Tsukamoto 2011; Liappas et al. 2007).

2.3. Cluster Analysis

Since primitive times, man felt the need to classify things according to their different characteristics, applying this categorization to his day-to-day life, from food classification, to organizing words into a language. Later these forms of categorization evolved into scientific and engineering methodologies and tools (Everitt et al. 2011).

The advances in technology, mainly personal computers has made multivariate analysis more accessible to data analysts, hence, widening its range of users and areas where it has been applied. The use of statistical software facilitates the analysis of large data sets, otherwise almost impossible to handle (Hair et al. 2007).

To better study large data sets, it is necessary to aggregate similar objects into smaller groups. By clustering similar individuals together they can be studied as a whole, making the data easier to manage (Everitt et al. 2011; Ward 1963). Human beings can also be classified and clustered together based on socio-economic status, behavior or habits (Everitt et al. 2011).

Romesburg (2004) wrote a simple yet interesting definition for cluster, "A cluster is a set of one or more objects that we are willing to call similar to each other. A cluster can be as few as one object, if we are willing to call no other objects similar to that object. Or it can be as many as all of the objects in the data matrix, if we are willing to call all of them similar to each other. It may seem strange to use the word "willing", but that is exactly the right word. To call two or

more objects similar, we must be willing to neglect some of the detail that makes them non-identical. We must be tolerant to some of their differences."

2.3.1. Hierarchical Cluster Analysis

Clusters can be Hierarchical or Non-Hierarchical. The most common non-hierarchical clustering method is designated K-Means, in short, it creates a pre-established number of clusters, using the nearest mean as a clustering criterion (Coussement et al. 2011). It starts with a pre-established number of clusters (*K*) formed by *K* objects as centroids , then assign the remaining objects to the cluster with the nearest centroid or mean distance (Arabie et al. 1996; Johnson & Wichern 2007).

Hierarchical cluster analysis (HCA) consists of a method where instead of grouping every cluster in just one step, it builds clusters by levels or partitions that go from a smaller cluster with only two objects, to a single cluster containing all objects, so we start with as many clusters as objects (Johnson & Wichern 2007). There are two types of hierarchical clusters, Agglomerative and Divisive. The first agglomerates the smaller clusters into bigger ones until it reaches a cluster with all individuals in it, the divisive method, in opposite, starts with a big cluster and divides it into smaller ones (Everitt et al. 2011; Kaufman & Rousseeuw 2009).

2.3.2. Data Matrix

The Data Matrix is the starting point for the cluster analysis. It shows the relationship between objects and attributes. In this case, the persons in the test are the objects, and the parameters such as age or B12 levels are their attributes. The cluster analysis tries to find which objects are similar, based on their attributes (Romesburg 2004).

A standardized Data Matrix can be calculated to remove the effect of different metric units in the data. These processes transform the values into dimensionless units making their contributions more unbiased. However, cluster analysis can run with either the data matrix or it's standardized version (Romesburg 2004).

2.3.3. Standardized Coefficients

When variables for the cluster analysis come in different scales, it is important to use standardization. This process converts the objects' values to dimensionless units, removing most discrepancies caused by variables in different units (Coussement et al. 2011).

In this study, the parameters come in different units, so standardization was necessary, therefore Z-Scores method was chosen.

Z-scores also referred as Standard Scores, is a simple method to standardize data that compares each value with the sample's mean and standard deviation (Mitrushina et al. 2005; Rubin 2012):

$$z - scores = \frac{X - \bar{X}}{\sigma} \tag{2.1}$$

Where X is the value being standardized, and $\bar{X}$ and σ are respectively the mean and standard deviation for the sample.

If a z-score is positive, it means its value is above the mean, and if it is negative, the value is below the mean.

2.3.4. Clustering Methods

In hierarchical clustering the most used methods are the linkage methods (single, complete and average), centroid method and variance or Ward's method (Malhotra 2006). These approaches will be described in this section.

The Single Linkage clustering method consists in finding the shortest distance between two objects, uniting them in one cluster. Then, the process runs again looking for the next shortest distance between two objects or clusters, this time the distance between an object and the cluster is the minimum distance between the object and either of the cluster's individuals (Romesburg 2004).

The Complete Linkage method, is the opposite of the Single Linkage in the way that it clusters objects based on the farthest distance between them (Malhotra 2006).

The Average Linkage process clusters objects using the average or mean distance between all pairs of objects in a cluster (Johnson & Wichern 2007; Malhotra 2006).

The Centroid method assumes the centroid distance as the distance between two clusters. Where a centroid is the cluster's center of mass (Romesburg 2004).

Ward proposed a method with particular focus on information loss. In this method, he tried to minimize the loss of information associated with the grouping of objects that are not necessarily or exactly similar. This loss would be visible in the Error Sum of Squares, obtained by equation 2.2 (Ward 1963):

$$SS_E = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \quad (2.2)$$

Where *n* is the number of objects and x_i is the score of the *i*th object.

If there are *k* groups or clusters, the SS_E is given by the sum of each group's SS_E:

$$SS_E = SS_{E_1} + SS_{E_2} + \cdots + SS_{E_k} \quad (2.3)$$

When uniting clusters, Ward's Method analyses all possible combinations and chooses the one that minimizes the SS_E. The method can be used by optimizing any objective function other than the error sum of squares (Ward 1963).

2.3.5. Resemblance Coefficient

The resemblance coefficient is the measure of the distance or proximity of two individuals, these are called respectively dissimilarity or similarity coefficients. In a dissimilarity coefficient, a smaller value (or distance), indicates a higher proximity. Contrary to this, a higher similarity value indicates a higher proximity (Romesburg 2004).

For this study, the Squared Euclidean Distance will be used as the resemblance coefficient for the cluster analysis.

Euclidean distance is the length of the straight line between two points. It is one of many methods to obtain a distance or dissimilarity coefficient between entities, it is calculated through the following equation (Johnson & Wichern 2007; Kaufman & Rousseeuw 2009):

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2} \quad (2.4)$$

Objects with similar distances between them, are agglomerated becoming a cluster. This value cannot be smaller than 0, the maximum similarity value (Romesburg 2004; Kaufman & Rousseeuw 2009).

The Squared Euclidean Distance, like the name suggests, is the square of equation 2.4. In practical terms, it ignores the root extraction effect (Pereira 1999). When using Ward's method it's advisable to use squared euclidean distances (Malhotra 2006).

2.3.6. Resemblance Matrix

With the results of the dissimilarity or similarity coefficient, it is possible to build a resemblance matrix with the distances or similarity between objects (Romesburg 2004).

Table 2.1 exemplifies a resemblance matrix with data from patients with respiratory disease in our study, this table is merely an example and no conclusions should be drawn from it. The matrix is square, showing the distance between each pair of objects. Furthermore, it is symmetrical since the distance from 1 to 2 is the same as from 2 to 1 (Romesburg 2004).

Table 2.1 - Resemblance matrix example

Case	Euclidean Distance						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	253.984	2218.996	1816.037	324.929	143.778	954.417
2	253.984	0	2242.036	1860.092	493.679	265.206	932.085
3	2218.996	2242.036	0	452.752	1955.348	2156.505	1345.143
4	1816.037	1860.092	452.752	0	1541.91	1761.73	1009.844
5	324.929	493.679	1955.348	1541.91	0	292.529	775.253
6	143.778	265.206	2156.505	1761.73	292.529	0	884.249
7	954.417	932.085	1345.143	1009.844	775.253	884.249	0

2.3.7. Dendogram

The Dendogram is a type of tree diagram, built with the values from the resemblance matrix, it exposes the clusters between objects for an easier and more intuitive analysis (Romesburg 2004).

In figure 2.1 an example of a dendogram is shown using the data from the resemblance matrix in table 2.1. The dendogram should be read from the left to the right, in divisive clustering, starting with the smaller clusters and moving on to the bigger ones until we get a single cluster (Malhotra 2006). It is possible to see each step of the clustering process, starting with the formation of the first cluster, uniting the two objects that are closer, 1 and 6. Then the smallest distance is between 1 and 2, since 1 is already clustered with 6, the second cluster is the combination of the cluster (1,6) and the object 2. The horizontal lines show the distance between the objects (Romesburg 2004; Pereira 1999).

Although it is not absolutely necessary, several resemblance matrixes can be calculated for each step of the dendogram, to better visualize the distances. This is especially helpful when handling large data sets (Abonyi & Feil 2007).

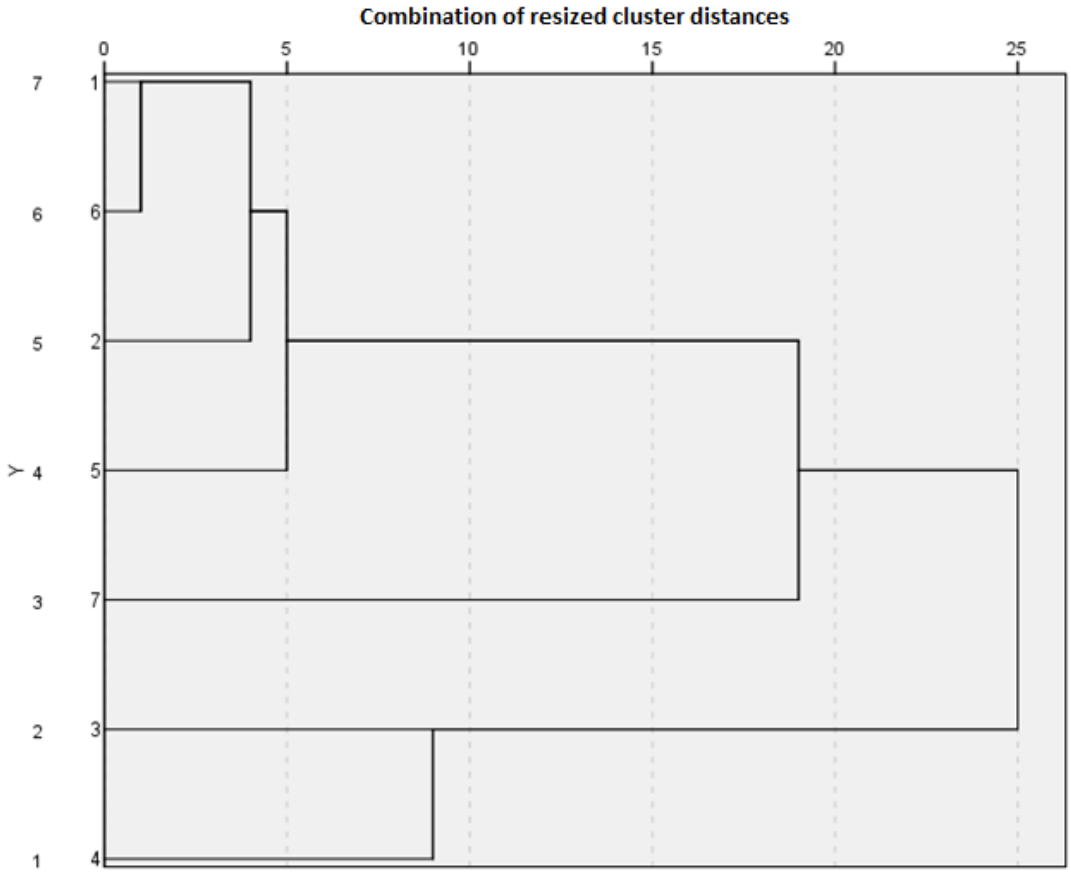

Figure 2.1 - Dendogram exemple

2.4. *One-Way Analysis of Variance*

Analysis of variance or ANOVA is a statistical tool used to test differences between group means. One-way ANOVA means that it only tests one independent variable. It tests the null hypothesis that the means of groups in that variable are equal or similar. It uses the F-test as a decision criterion (Westfall et al. 1999; De Muth 2006).

To submit data to an ANOVA test, one must make sure it complies with its assumptions of homoscedasticity (similar variances) and that the dependent or predictor variable must be normally distributed (De Muth 2006).

For *k* population means, the hypothesis test would be (Tamhane 2009):

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: \mu_i \neq \mu_j \end{cases} \quad (2.5)$$

Where, μ is the group's mean.

The F-test is then conducted with the Sum of Squares method. Following the explanation in the bibliography (Montgomery & Runger 2010; Tamhane 2009), we arrive at equation 2.6:

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2=\sum_{i=1}^{a}n_i(\bar{y}_i-\bar{\bar{y}})^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 \tag{2.6}$$

Which can be simplified to:

$$SS_T=SS_B+SS_W \tag{2.7}$$

Where SS_W stands for Within Sum of Squares, and is a direct reflection of the variation among individuals in the same conditions, SS_B is the Between Sum of Squares, besides the inherent variation for individuals in the same conditions it also reflects any effect from differential treatment. SS_T is the Total Sum of Squares (Coladarci et al. 2010; Montgomery & Runger 2010).

To calculate the critical *F*-test value, it's necessary to calculate the Mean Squares for both the between and within sum of squares, as shown in equation 2.8 (Tamhane 2009):

$$F_0=\frac{MS_B}{MS_W}=\frac{SS_B/\vartheta_B}{SS_W/\vartheta_W} \tag{2.8}$$

Where ϑ is the number of degrees of freedom in the between or within sources of variation. The critical *F* value is then compared with the test *F* and H_0 is rejected if $F_0 > f_{\alpha,\vartheta_B,\vartheta_W}$, where α is the chosen significance level.

To better access and visualize the above information it is usually compiled in the ANOVA table, as shown in table 2.2 (Montgomery & Runger 2010):

Table 2.2 - ANOVA general form

Source of variation	**Sum of Squares**	**d.f**	**Mean squares**	***F***
Between	SS_B	$k-1$	MS_B	F_0
Within	SS_W	$n-k-1$	MS_W	
Total	SS_T	$n-1$		

Where *n* is the number of observations and *k* is the number of means or groups.

2.5. Nonparametric Analysis

Parametric statistic tests make certain assumptions about data, most commonly that it is normally distributed or that the observations are independent from each other. It is also an assumption of normality that the dependent variable is in a numeric scale form and the population consists of at least 30 individuals. Seldom real data matches all of a test's assumptions, and these restrictions in assumptions limit parametric tests. Although some tests are robust enough to condone the violation of certain assumptions, others simply cannot (Kvam & Vidakovic 2007; Ferraty & Vieu 2006).

Nonparametric statistics come to play when a parametric test's assumptions are not matched by the data. Research shows that parametric statistics are more powerful than their nonparametric counterpart, only if all assumptions are met. Nonparametric statistics are able to work with smaller samples, adapt to more irregular distributions, and mostly use ordinal data rather than scales (Kvam & Vidakovic 2007; Ferraty & Vieu 2006; Pett 1997).

Nonparametric tests' versatility makes them ideal for health care research (Pett 1997).

2.5.1. Kruskal-Wallis One-Way Analysis of Variance by Ranks

The Kruskal-Wallis one-way analysis of variance by ranks (Kruskal & Wallis 1952) is a non-parametric alternative to the one-way ANOVA. It is used when the data does not comply with the assumptions of the parametric ANOVA.

Kruskal-Wallis test uses rank order (ordinal) data, and indicates if there is a significant difference between the medians of two or more independent groups. When converting from ratio or interval data to rank data, some of the information is lost, making this test slightly less powerful than its parametric equivalent, the one-way ANOVA (Sheskin 2003; Plichta & Garzon 2009).

The test begins, analogously to the ANOVA, with a hypothesis test (King et al. 2011):

$$\begin{cases} H_0: M_1 = M_2 = \cdots = M_k \\ H_1: M_i \neq M_j \end{cases} \tag{2.9}$$

Where *M* is the median for each group (Osborn 2000).

Then we calculate the Kruskal-Wallis's test statistic, *H*:

$$H = -3(N+1) + \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \tag{2.10}$$

Where *N* is the total number of individuals, *n* is the number of individuals in each group and *R* is each individual's rank (King et al. 2011).

The test *H* value is then compared with the tabulated value of the chi-square distribution, $\chi^2_{\alpha;\vartheta}$, where α is the significance level and $\vartheta = k - 1$ is the number of degrees of freedom. *k* is the number of groups. If $H > \chi^2_{\alpha;\vartheta}$ the null hypothesis is rejected and there are significant differences among the groups (King et al. 2011).

2.5.2. Spearman Correlation

Spearman correlation is most commonly used as the nonparametric alternative to the Pearson correlation, and can be interpreted as a special case of the Pearson correlation where ordinal data is used instead of scale data, and does not follow normal distribution patterns (Weinberg & Abramowitz 2002; Gravetter & Wallnau 2013). However when the data is normally distributed, Pearson correlation is a more powerful tool (O'Rourke et al. 2005).

Spearman Correlation (ρ_S) is a nonparametric statistic that can be used to find relationships between two variables, using ordinal data. It can measure the degree of linear correlation among two variables, for example to which extent *y* increases or decreases when *x* increases or decreases, thus being a bivariate correlation statistic (Gravetter & Wallnau 2013).

To perform a Spearman correlation, since it works only with ordinal or rank variables, one must first convert scale data to ordinal, for each variable separately, for example *x* and *y*. Then calculate the sum of squares of the differences *d* between the *x* and *y* ranks. It is now possible to calculate Spearman's rho through equation 2.11 (Deep 2006):

$$\rho_s = \frac{6\sum R_i^2}{n(n^2-1)} \tag{2.11}$$

Where *n* is the number of pairs of values. When $n > 10$, the unit normal equation 2.12 is used:

$$Z = \frac{n(n^2-1) - 6\sum R_i^2}{n(n+1)\sqrt{n-1}} \tag{2.12}$$

The range of the coefficient is $-1 \leq \rho_S \leq 1$, where 1 and -1 are a perfect correlation, -1 indicates a negative correlation and 0 indicates there is no correlation at all (Sheskin 2003; Deep 2006).

3. Methodology

3.1. Clinical Analysis Data

The data from came from a Portuguese clinical analysis laboratory, and it consists of blood analysis required by the patients' doctors or from routine checks and were retrieved in 2011.

This is a population that was seeking medical attention, and was directed to the clinical laboratory for diagnosis.

The total population consisted of 728 individual collectings, which correspond to the samples that include the six desired parameters, folic acid, vitamin B12, ferritin, transferrin, serum iron and hemoglobin, from persons residing in Lisbon and Setubal's metropolitan area. The research described in section 2.2.4 led us to remove the pregnant women from the study, to avoid the risk of contaminating the data.

When coming in for examination, the patients were asked routine questions about their age, gender, county of residence, why they came for examination and what medication or supplementation they were taking.

The population consisted of 140 male patients and 588 female patients, aged from 2 to 95 years of age. There are 329 persons over 60 years old, these are considered elderly patients.

With the questions about why they came in for examination and what medication or supplementation they were taking, it was possible to divide them into 14 categories:

- Allergy;
- Anemia: this group contains patients with anemia, and patients whose values and supplementation history suggesting anemia;

- Anti-Inflammatory: persons taking anti-inflammatory medication;
- Diabetes;
- Autoimmune Disease: this group contains patients with systemic lupus erythematosus, and unspecified autoimmune diseases;
- Cardiovascular Disease: group containing patients with hypertension, hypercholesterolemia, taking anticoagulant medication and specific medication indicating cardiovascular diseases;
- Gastrointestinal Disease: this group consists of persons with gastric ulcers, Crohn's disease, weight loss process, anorexia, taking proton bomb inhibitors, antacids, interventions for gastric banding, gastric bypass, gastrectomy and coeliac and gastric diseases;
- Liver Disease: patients with hepatitis B, bladder problems and specific medication indicating liver disease;
- Neurological Disease: consists of dementia, Parkinson, epilepsy, Alzheimer, depression and anxiety;
- Renal Disease: dialysis patients and specific medication indicating renal diseases;
- Respiratory Disease: patients with asthma and specific medication indicating respiratory disease;
- Rheumatic Disease: persons suffering from rheumatoid arthritis, ankylosing spondylitis, taking salazopyrin and specific medication indicating rheumatic diseases;
- Neoplasia;
- Routine: patients coming to the clinic for routine consultations.

There are 310 routine patients. These patients were removed from the study, before the actual multivariate analysis, based on the lack of factual information about these patients since most of them were being tested for possible diseases that they were not aware of and so did not report them to the analyst. We felt this lack of information would cloud the results leading to poor conclusions.

The anonymity of the persons involved in the study was kept by attributing each individual a codename consisting of seven digits defined by the clinical laboratory's software.

To assess the information from a population's analysis, the parameter's levels must be compared with reference levels and the population's levels (Burtis et al. 2012).

Reference levels for folic acid, B12, ferritin, transferrin iron and hemoglobin used in the laboratory where the samples were taken are described in table 3.1.

Table 3.1 - Reference values for biomarkers

	Man	Women
Folate	>3.3 ng/mL	>3.3 ng/mL
Vitamin B12	>273 pg/mL	>273 pg/mL
Ferritin	30 – 400 µg/L	10 – 270 µg/L
Transferrin	200 – 360 ng/dL	200 – 360 ng/dL
Iron	59 – 158 µg/dL	37 – 145 µg/dL
Hemoglobin	13 – 17 g/dL	12 – 16 g/dL

With little consensus when it comes to reference values (Refsum et al. 2004), these are the most commonly accepted in the bibliography (Almeida et al. 2012; Burtis et al. 2012; Clarke et al. 2004; Clarke et al. 2003; Gibson 2005; Herbert 1987; Refsum et al. 2004).

Also according to the literature (Herbert 1987; Gibson 2005), folate levels of 3.3 ng/mL are extremely low (negative or depletion status). This value would only make sense if we were studying extremely severe cases. Based on the same literature, folate is considered low when below 5 ng/mL, and this will be considered for evaluation of individuals in this study.

3.2. Descriptive Statistics

The first step towards the characterization of this population was obtaining descriptive statistics such as means, medians, standard deviations and the interquartile range (25^{th} and 75^{th} percentiles) for the parameters in the study. These were obtained recurring to Microsoft Excel (MS Excel). This process will help assess the condition of the population in general.

In this step the descriptive statistics were divided into two sets, one with means and standard deviations and the other with percentiles and interquartile ranges. The second process became necessary after calculating the mean and standard deviation for the data and finding that some of the variables had great standard deviations, in some cases even greater than the respective means. Calculating the medians and percentiles provided another view of the data so conclusions could be drawn based on more and more accurate information.

3.3. Spearman Correlation

Iron, ferritin and transferrin are interrelated (Burtis et al. 2012), so a Spearman correlation test was conducted to verify if they are statistically related. If so, it is necessary to remove dependent variables, otherwise we are at risk of retrieving poor results from a biased statistical test.

3.4. One-Way ANOVA

According to Andrès et al. (2007), Baik (1999) and Campbell et al. (2003) folate absorption decreases with age. To find statistical evidence of this phenomenon the population was separated into four age groups. The first group consists of persons aged from 0 to 29 years old, the second from 30 to 59 years old, the third group is composed of persons from 60 to 79 years old and finally the last group is for persons of 80 years and older. The individuals in the last two groups are considered elderly citizens. One-way ANOVA was used to see if they were significantly different, based on their folate, cobalamin, iron and hemoglobin levels. One-way ANOVA assumptions were tested with resource to Harley, Cochran and Bartlett's tests and the data does not comply with the assumptions. Thus it was necessary to perform an equivalent non-parametric test, the Kruskal-Wallis *H* test.

3.5. Cluster Analysis

After the decision was made about what data was going to be left out, the dataset was submitted to a Cluster Analysis using IBM SPSS Statistics 22.

Ward's clustering method was chosen since it is indicated for large data sets (Ward 1963), with the squared euclidean distances.

Data was standardized using the Z-scores method, thus removing the effect of metric units.

The cut made to define the number of clusters was made with folate as the main indicator, but never neglecting the remaining variables.

Tables were then drawn with the means and standard deviations for the parameters in study (Folate, Vitamin B12, Iron and Hemoglobin), for each cluster and also with the number of illnesses by cluster, so the data could be cross referenced and we could assess what illnesses were affected by what variables. In this process MS Excel was the chosen tool.

To deepen our knowledge in the clusters and try to relate diseases with the biomarkers, we used the filters in MS Excel to find in each cluster, which diseases registered parameter concentrations bellow the reference values.

3.6. Sensitivity Analysis

To test the robustness of the cluster analysis, a Sensitivity Analysis was performed.

This sensitivity analysis consisted of performing a cluster analysis, merging together the age groups that are not significantly different.

Similarly to the previous cluster analysis Ward's method was used with the squared euclidean distances and data was standardized with z-scores.

4. Results and Discussion

4.1. Descriptive Statistics

The first step of a statistic study is to describe the population, for this purpose basic descriptive statistics are used, such as means and standard deviations (st dev). Table 4.1 shows those statistics for our population.

Table 4.1 - Means and standard deviations for the population

Age		0-29	30-59	60-79	>80	Total
Folate (ng/mL)	Mean	8.09	8.79	10.97	12.15	9.86
	St Dev	4.83	4.73	6.16	6.85	5.74
B12 (pg/mL)	Mean	483.30	448.64	521.79	583.98	496.47
	St Dev	234.39	244.19	372.89	470.36	331.18
Hemoglobin (g/dL)	Mean	12.80	12.41	12.04	11.88	12.27
	St Dev	1.47	1.60	1.71	1.75	1.66
Iron (µg/dL)	Mean	82.40	81.47	71.65	70.03	76.91
	St Dev	42.88	45.19	35.63	32.76	40.68
Ferritin (µg/L)	Mean	63.77	76.03	115.63	172.54	101.16
	St Dev	94.63	148.85	139.53	259.06	166.67
Transferrin (ng/dL)	Mean	295.00	291.42	262.95	241.88	275.74
	St Dev	64.79	59.65	54.11	59.11	61.82

Table 4.1 shows that for ages from 0 to 29, folate levels are the lowest of the population means, which can be an indicator of poor folate ingestion at young ages (Taneja et al. 2007).

Elderly citizens however show a higher folate means than the population's total. These values are above the 5 ng/mL indicator.

B12 levels are good in general for all age groups, as it should be above 273 pg/mL. The standard deviations however are very high when compared to the mean values, which is an indicator of a wide range of results that need a closer look.

Hemoglobin levels are all extremely close to the reference value of <12 g/dL for all age groups, this reveals a general low hemoglobin for the population.

Iron means are within the normal range (37 – 145 µg/dL in women and 59 – 158 µg/dL in men), in this parameter we must also be careful to look for values above the reference range since high values for iron are as hazardous as low values. Analogously to the B12 vitamin, the standard deviations for this parameter are high in comparison with the means, so further study is necessary to achieve better results.

Ferritin mean values all fit the reference intervals (10 – 270 µg/L in women and 30 – 400 µg/L in men), although they are lower in the first two age groups and increase with age. However the standard deviation values for this variable are very high and make it hard to assess the real state of ferritin in the population.

Transferrin values are all normal, within the limits of 200 – 360 ng/dL.

This data has a high variation problem that reflects in the standard deviation values, these high standard deviations might not correctly represent the whole range of results. To get another look at the data and remove any bias that can come from the means and standard deviations format, it is possible to use medians and interquartile ranges for 25% and 75% percentiles. Results are shown in table 4.2. There it becomes easier to perceive the range of the data. Median values for folate are lower than the average, which means there are more low values for folate than high ones. The same behavior is visible for B12 and iron values. Hemoglobin however shows median values close to the means, this is expected if we look at the low values for the standard deviation in table 4.1, which indicate a narrower range of values for this variable.

Despite the relatively high standard deviation values shows similar results for means and medians. This indicates that there are approximately as many low values as high values. Ferritin median values are much lower than the means. This also means that there are more low values of ferritin than high values, however the variance associated with this variable is extremely high.

There is also an increase in ferritin values in elderly citizens that exhibit median values of more than twice the value from under 60 years of age.

The data in tables 4.1 and 4.2 still contains data from routine checks, since their purpose is characterizing the population and we are not looking for correlations at this point.

Table 4.2 - Medians and interquatile ranges of 25% and 75% for the population

	Age		0-29	30-59	60-79	>80	Total
Folate (ng/mL)	Median		6.55	7.30	9.45	10.60	8.10
	Percentile	25%	4.85	5.80	6.63	6.55	5.90
		75%	10.13	10.05	13.48	15.85	11.90
B12 (pg/mL)	Median		416.50	395.00	403.00	419.00	399.00
	Percentile	25%	329.00	313.50	315.50	335.00	319.00
		75%	591.75	507.00	565.25	589.00	543.00
Hemoglobin (g/dL)	Median		12.80	12.50	12.20	11.80	12.40
	Percentile	25%	12.10	11.60	11.20	10.65	11.30
		75%	13.63	13.40	13.10	12.80	13.40
Iron (µg/dL)	Median		76.50	79.00	72.00	69.00	74.00
	Percentile	25%	49.50	46.00	44.00	43.50	45.00
		75%	106.25	109.00	94.00	90.50	102.00
Ferritin (µg/L)	Median		30.60	32.70	68.40	88.90	46.65
	Percentile	25%	13.05	10.30	25.38	39.10	17.70
		75%	81.88	75.50	141.23	196.90	109.13
Transferrin (ng/dL)	Median		290.00	282.00	257.00	239.00	266.00
	Percentile	25%	246.25	246.50	227.00	208.00	232.00
		75%	339.50	330.00	289.75	279.00	316.00

Another phenomenon we want to study is how the values for these parameters relate to the patient's health condition. For that table 4.3 was built.

From table 4.3 we can see that persons who suffer from allergies and liver disease are the ones with lower folate levels. But if we take a closer look it is visible that in most illnesses the lower percentiles are also low, indicating the same tendency for low folate levels for the population in general.

Vitamin B12 levels are normal in all cases, and even lower percentiles are far from the reference values of low cobalamin levels.

Table 4.3 - How parameter levels relate to patient's condition

	Nr. Persons			Folate	B12	Hemoglobin	Iron	Ferritin	Transferrin
Allergy	9	Median		6.20	559.00	13.10	101.00	38.30	278.00
		Percentile	25%	5.40	320.00	12.70	69.00	30.70	248.00
			75%	10.10	711.00	14.80	105.00	112.20	284.00
Anemia	80	Median		8.55	407.00	11.70	54.50	29.60	296.00
		Percentile	25%	6.10	319.75	10.40	28.50	7.08	241.50
			75%	15.23	595.00	12.80	90.25	74.53	356.00
Anti-inflammatory	3	Median		9.60	370.00	11.90	93.00	44.20	286.00
		Percentile	25%	9.00	302.00	11.35	55.00	34.30	259.00
			75%	15.05	452.50	12.80	150.00	80.30	306.50
Diabetes	84	Median		8.90	390.50	11.90	60.00	50.95	266.00
		Percentile	25%	6.30	310.00	10.78	38.50	21.23	232.75
			75%	12.90	571.50	12.90	78.00	128.30	302.25
Autoimmune Disease	5	Median		9.50	371.00	12.70	44.00	9.40	273.00
		Percentile	25%	6.30	327.00	12.10	43.00	7.90	254.00
			75%	13.88	511.00	13.50	46.00	28.40	277.00
Cardiovascular Disease	111	Median		9.20	397.00	12.40	76.00	66.60	254.00
		Percentile	25%	6.40	323.50	11.20	57.00	35.50	223.00
			75%	12.05	479.00	13.60	94.00	135.05	286.50
Gastrointestinal Disease	41	Median		8.40	397.00	12.90	83.00	50.80	257.00
		Percentile	25%	6.10	314.00	11.90	46.00	20.80	227.00
			75%	11.20	509.00	14.20	130.00	103.60	303.00
Liver Disease	3	Median		6.00	490.00	12.30	107.00	318.20	220.00
		Percentile	25%	4.65	365.50	12.30	84.00	201.95	205.00
			75%	12.30	549.50	13.70	109.00	423.55	234.50
Neurological Disease	34	Median		7.45	387.00	12.65	70.50	38.25	272.00
		Percentile	25%	5.20	315.25	11.60	42.25	12.25	233.00
			75%	8.38	529.50	13.98	100.75	62.68	329.25
Renal Disease	3	Median		18.00	448.00	11.70	101.00	124.70	216.00
		Percentile	25%	10.75	420.00	10.70	88.00	73.45	209.00
			75%	18.90	457.50	12.15	107.00	456.75	260.50
Respiratory Disease	7	Median		9.40	563.00	12.40	59.00	34.20	276.00
		Percentile	25%	6.55	396.50	11.50	57.50	9.40	246.50
			75%	13.05	610.50	13.45	82.50	140.50	348.00
Rheumatic Disease	11	Median		8.10	384.00	11.70	91.00	49.70	261.00
		Percentile	25%	6.25	362.00	11.20	76.00	23.10	247.50
			75%	15.60	1034.50	12.20	113.50	88.15	289.50
Neoplasia	4	Median		11.65	698.50	12.00	72.00	43.40	309.00
		Percentile	25%	8.03	567.00	10.53	46.75	16.58	257.50
			75%	17.35	841.25	12.83	100.25	104.35	358.75
Thyroid	25	Median		7.60	363.00	12.50	82.00	22.70	287.00
		Percentile	25%	6.65	294.50	12.00	64.00	2.90	239.50
			75%	12.65	462.50	13.10	103.00	59.35	326.50

The tendency for low levels of hemoglobin also shows in this table, with some of the median values below the 12 g/dL reference value, these are anemia, anti-inflammatory, diabetes, renal disease and rheumatic disease.

As for iron, anemia, autoimmune disease and respiratory disease. Even if these values are above the reference value, they are indeed low and might show a correlation between low iron and these conditions.

Ferritin is a good indicator for iron levels, so it is no surprise to see low ferritin levels for the illnesses with low iron. However, allergy, neurological disease and especially thyroid also show low levels of ferritin.

Transferrin levels are all within the reference values and show no patterns that relate high or low levels with high or low iron levels.

This first analysis of means and medians gives us a general idea of the population we are studying. Another interesting statistic is the number and percentage of individuals with parameter concentrations under the reference values in table 3.1. Such data could be assessed in table 4.4.

Table 4.4 - Number and percentage of persons with levels under the reference

	0-29 (104)	30-59 (295)	60-79 (214)	>80 (115)	Total	%
Folate	28	35	24	11	98	13.5
B12	10	46	28	14	97	13.3
Iron (low)	15	57	38	21	131	18
Iron (high)	9	26	6	1	42	5.8
Hemoglobin	23	101	25	63	283	38.9
Ferritin	23	73	13	7	121	16.6
Transferrin	4	9	16	21	50	6.9

() is the number of persons in that age group.

Table 4.4 shows the number of persons suffering from levels below the reference values, and in the case of iron, also values over the reference. Also percentage of the total population that they represent.

The percentage of individuals with folate values under 5 ng/ml is 13.5%. The age group between 30 and 59 years of age is the one that contributes the most for this percentage. The

first age group, which corresponds to the young persons is the one that exhibits more individuals with folate deficiency compared to the number of persons in the group.

Vitamin B12's values are similar to folate's (13.3). Once more the second age group is the one with most deficiency levels.

Low serum iron affects 18% of our study population. The second and third age groups are the ones that register more deficient values. Iron values above the reference values in table 3.1 are the ones with the least percentage, only 5.8% of the total population. Analogously to the previous parameters the second group is the most affected.

As tables 4.1 and 4.3 suggest, hemoglobin concentrations in the population are generally low. Table 4.4 helps us see that 38.9% of the total population is hemoglobin deficient. Persons aged 30 to 59 years are once more the most affected group, although persons over 80 years have a higher number when compared to the total number of individuals in the age group.

Being an indicator for poor serum iron concentrations, ferritin shows a similar percentage to iron, 16,6%.

Low transferrin affects 6.9% of the population. It is interesting to see how transferrin deficiency increases with age.

Table 4.5 shows a similar study to table 4.4, with the exception that the routine patients were removed, and will be removed from the study henceforth, to better assess the effect of the biomarkers in the patient's condition.

Table 4.5 - Number and percentage of persons with levels under the reference (without routines)

	0-29 (34)	30-59 (152)	60-79 (157)	>80 (75)	Total	%
Folate	12	12	17	7	48	11.5
B12	5	25	19	9	58	13.9
Iron (low)	9	32	30	12	83	19.9
Iron (high)	3	14	3	1	21	5.0
Hemoglobin	11	50	18	42	179	42.8
Ferritin	12	39	12	3	66	15.8
Transferrin	1	4	13	12	30	7.2

() is the number of persons in that age group.

As it was expectable, results in table 4.5 do not vary much from table 4.4. However it is possible to see that the percentage of low hemoglobin and iron has increased, suggesting that these parameters affect or are affected by the illnesses in the study.

To better understand the characteristics of the population and how the biomarkers relate among themselves and with the patient's illnesses, more elaborate statistical tests are necessary.

4.2. Normality Tests

Before introducing the variables in any statistic test, a standard normality test was conducted with IBM SPSS Statistics 22, using the Kolmogorov-Smirnov goodness of fit test, and Shapiro Wilks test. The statistical tests that verify if the data fits a normal distribution. The results are shown in table 4.6.

Table 4.6 - Normality tests

	Kolmogorov-Smirnov			**Shapiro-Wilk**		
	Statistic	d.f.	Sig.	Statistic	d.f.	Sig.
Folate	.092	418	.000	.947	418	.000
Ferritin	.286	418	.000	.517	418	.000
B12	.203	418	.000	.683	418	.000
Iron	.064	418	.000	.942	418	.000
Transferrin	.059	418	.001	.983	418	.000
Hemoglobin	.211	418	.000	.761	418	.000

As it is visible in table 4.6 none of the variables in the study follow a normal distribution, so it's safe to say we're dealing with non-parametric data. This type of data requires a careful approach since normality is an assumption for most statistics.

4.3. Spearman Correlation

The Spearman Correlation test was also carried with IBM SPSS Statistics 22. The objective of this test is to verify statistically that iron, ferritin and transferrin are, as literature suggests, interdependent. The test results are in table 4.7. There we can see that iron, ferritin and transferrin are correlated for a 0.01 significance level, which means they are dependent variables.

Table 4.7 - Spearman correlation test

			Folate	Ferritin	B12	Iron	Transferrin	Hemoglobin
Spearman's rho	**Folate**	Correlation Coeficient	1.000	-.003	.056	-.017	.082	.026
		Sig. (2 way)		.956	.250	.727	.095	.590
		N	418	418	418	418	418	418
	Ferritin	Correlation Coeficient	-.003	1.000	.146**	.423**	-.666**	.188**
		Sig. (2 way)	.956		.003	.000	.000	.000
		N	418	418	418	418	418	418
	B12	Correlation Coeficient	.056	.146**	1.000	.025	-.117*	-.042
		Sig. (2 way)	.250	.003		.616	.017	.393
		N	418	418	418	418	418	418
	Iron	Correlation Coeficient	-.017	.423**	.025	1.000	-.262**	.430**
		Sig. (2 way)	.727	.000	.616		.000	.000
		N	418	418	418	418	418	418
	Transferrin	Correlation Coeficient	.082	-.666**	-.117*	-.262**	1.000	-.131**
		Sig. (2 way)	.095	.000	.017	.000		.007
		N	418	418	418	418	418	418
	Hemoglobin	Correlation Coeficient	.026	.188**	-.042	.430**	-.131**	1.000
		Sig. (2 way)	.590	.000	.393	.000	.007	
		N	418	418	418	418	418	418

* shows significative correlations for 0.05 and ** shows significative correlations for 0.01 significance.

Hemoglobin is also highly correlated with iron. A great part of hemoglobin is synthetized by transferrin bound iron (Roddie & Davis 2009). Although they are related hemoglobin can only be used as an indicator for iron when compared with the globular volume (mean corpuscular volume, MCV) (Semmelrock et al. 2012).

There is also a correlation between ferritin and vitamin B12.

Dependent variables cannot be inserted into the same model as independent so two need to be eliminated. As iron is a better indicator for serum iron levels, than transferrin and ferritin exhibits a great variance making it hard to fit into statistical models, transferrin and ferritin will be eliminated from this point on, leaving folate, vitamin B12, iron and hemoglobin as our four variables for this study.

4.4. One-Way ANOVA and Age Groups

To better study this population it was necessary to divide it by age groups. Based on the bibliographical research four groups were created, the first group consists of persons from 0 to 29 years of age, the second group contains persons from 30 to 59 years, the third from 60 to 79 and the last group contains those older than 80 years of age. The last two groups are considered elderly citizens.

For the groups to be statistically accurate it is necessary that they are, in fact, statistically different from each other. To test this hypothesis Statsoft Statistica software was used to perform a One-Way ANOVA test. Keeping in mind that the decisive parameter is folate, only folate results will be presented in the text as the others are provided in appendix A.

Table 4.8 - One-Way ANOVA for folate

Effect	SS	d.f.	MS	F	*p*-value
Intercept	28713.19	1	28713.19	869.35	0.000
Age	807.06	3	269.02	8.14	0.000
Error	13673.66	414	33.03		

As table 4.8 shows, the *p*-values are under the 0.05 significance, meaning the groups are in fact significantly different. As the data are non-parametric it is prudent to verify the ANOVA assumptions before proceeding with the study, mainly the variable's homoscedasticity:

Table 4.9 - Tests for ANOVA variance assumptions for folate

	Hartley F-max	Cochran C	Bartlett Chi-sqr	d.f.	*p*-value
Folate	2.748	0.365	14.277	3	0.002

Table 4.9 shows the results for Hartley, Cochran and Bartlett tests, and as we can see, the *p*-value for this test is under the 0.05 significance meaning they violate the variance assumptions for ANOVA.

When the assumptions for the parametric ANOVA are violated, one can perform its non-parametric equivalent, the Kruskal-Wallis *H* test. This test uses ranks and measures the difference between medians.

Table 4.10 - Kruskal-Wallis medians test for folate

	Age Group	1	2	3	4	Total
Median:	Observed	22.00	92.00	68.00	28.00	210.00
	Expected	17.08	76.36	78.88	37.68	
	Obs.-Exp.	4.92	15.64	-10.88	-9.68	
Median:	Observed	12.00	60.00	89.00	47.00	208.00
	Expected	16.92	75.64	78.12	37.32	
	Obs.-Exp.	-4.92	-15.64	10.88	9.68	
Total	Observed	34.00	152.00	157.00	75.00	418.00

For table 4.10 the $\chi^2_{0.05;3} = 17.2911$ and the *p*-value is 0.0006. Since *p*-value < 0.05, there are at least two groups with significant differences between them. Table 4.11 shows which groups are different, presenting the corresponding *p*-value.

Table 4.11 - Kruskal-Wallis significantly different groups for folate (*p*-values)

Age group	1	2	3	4
1		0.138	**0.001**	**0.000**
2	0.138		**0.037**	**0.006**
3	**0.001**	**0.037**		1.000
4	**0.000**	**0.006**	1.000	

Table 4.11 shows the groups that are significantly different, in this case with a significance value of 0.05, thus a *p*-value beneath 0.05 indicates that there are significant differences between two groups. The *H* value for this test is 26.4096. It is possible to see that group 1 is different from groups 3 and 4, group 2 is also different from groups 3 and 4. Groups 3 and 4 are not significantly different. Since this table is symmetric, the opposite is also true.

For a more intuitive view of the differences between these groups, it's possible to turn to the graphic view in figure 4.1.

Another thing that can be noticed both in figure 4.1 and the ones in appendix A is the fact that the first age group shows a great variance compared to the other groups. This group is associated with some uncertainty for several reasons. The fact that most of the study population is of the feminine gender and at this age it is common for women to take supplements (Hilton 2007; Shuaibi et al. 2008). Also at younger ages, folate levels are not yet stable and tend to grow with age (Kerr et al. 2009). At young ages folate levels are highly dependent of folate intake

(Taneja et al. 2007), information that is not accessible for this study. Finally the main focus of the study is the adult population and the group ranging from 0 to 29 years has a small number of individuals (only 34) compared with the others. All these factors led to the exclusion of this age group avoiding inconclusive and misleading results.

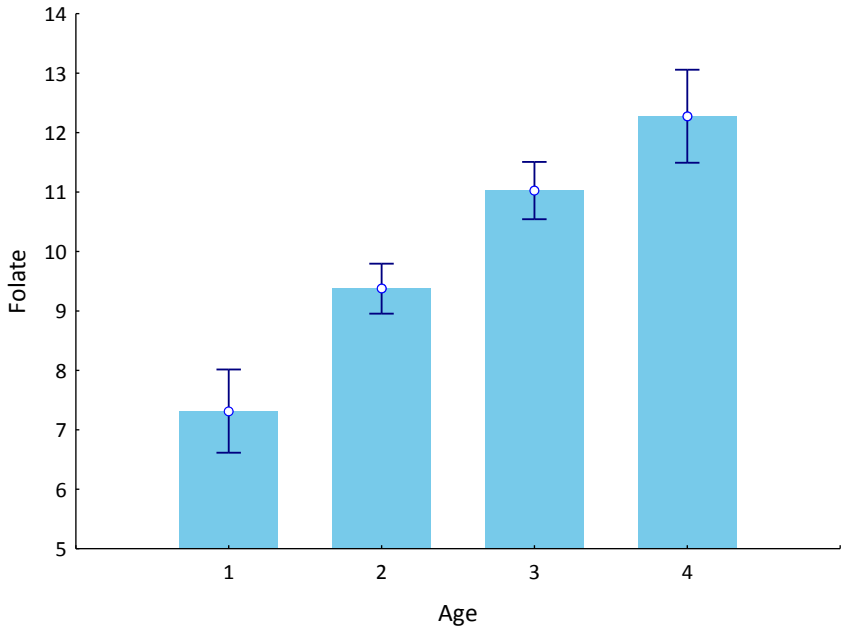

Figure 4.1 – Folate levels for each age group (the bars represent the means and the whiskers the standard deviations)

4.5. Cluster Analysis

Data was submitted to a cluster analysis for a more detailed view over the age groups. Relevant information is hidden in large data sets and cluster analysis can help dissect those large groups into smaller and more manageable ones. The cluster analysis was conducted with IBM SPSS Statistics 22 software.

A cluster analysis was performed for each age group and two tables were built for each analysis. One consists of the means and standard deviations for each cluster and the total population, the other indicates how many individuals are in each cluster and their respective illnesses. It is important to cross information between the two tables to look for possible correlations.

Due to restrictions in size, dendograms for each cluster must be displayed in appendix B instead of the main text.

4.5.1. Ages from 30 to 59 years

The first age group ranges from 30 to 59 years of age. This is the adult population group. The dendogram for this age group is displayed in appendix B, figure B.1.

Four clusters were formed in this group and the results can be consulted in tables 4.12 and 4.13.

Table 4.12 - Cluster analysis results for 30 - 59 years

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Folate (ng/mL)	Mean	7.72	12.03	18.20	7.78	9.37
	St Dev	3.38	8.94	5.49	2.65	5.19
B12 (pg/mL)	Mean	442.95	1698.00	470.27	391.75	456.69
	St Dev	181.36	223.26	152.99	127.76	258.14
Iron (µg/dL)	Mean	123.77	84.50	71.68	44.38	79.16
	St Dev	38.48	35.33	30.28	22.32	47.31
Hemoglobin (g/dL)	Mean	13.50	13.00	12.71	11.37	12.41
	St Dev	1.31	1.22	1.30	1.43	1.67

Table 4.13 - Number of persons with each illness aged 30 - 59 years

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Total	57	4	22	69	152
Allergy	2			2	4
Anemia	7		6	27	40
Anti-Inflammatory			1		1
Diabetes	4		2	11	17
Autoimmune Disease	1		2	2	5
Cardiovascular Disease	15	1	3	6	25
Gastrointestinal Disease	10		5	4	19
Liver Disease	1				1
Neurological Disease	6	2		7	15
Respiratory Disease	1		1	1	3
Rheumatic Disease	1	1		3	5
Neoplasia	1			1	2
Thyroid	8		2	5	15

Cluster 1

From this first cluster, illustrated in tables 4.12 and 4.13 we can see that folate levels are below the population average. Iron is above the population average, but not over the reference value (table 3.1). This cluster also has a high number of cardiovascular, gastrointestinal patients, and we can speculate that these justify the low folate and high iron levels. Vitamin B12 and hemoglobin levels in accordance with the reference values.

Using Microsoft Excel's filter options, it is possible to see, in each cluster, the biomarker concentrations that are under or over the reference values, and see which illnesses affect those patients. In this cluster there were seven patients with folate levels under 5 ng/mL, two with gastrointestinal disease, two with neurological disease (depression), one cardiovascular patient, one with allergy and one with respiratory disease.

We also found ten cases of cobalamin concentrations below 273 pg/mL, three with cardiovascular disease, two with gastrointestinal disease, two cases of anemia, one diabetes patient, one with thyroid complications and one with respiratory disease.

As the mean iron value suggests no serum iron cases below the reference value were registered, however high iron cases (above 145 µg/dL) are also dangerous as described in the literature review (chapter 2.1.3). In this cluster fourteen cases of high iron were found, these were observed in four patients with gastrointestinal disease, three with anemia, two neurological patients, one with allergy, one with diabetes, one cardiovascular patient, one with anemia and one thyroid patient.

There were also found four individuals with low hemoglobin levels, two with anemia, and one each with cardiovascular and gastrointestinal disease.

Cluster 2

The second cluster, has only four individuals with folate, iron and hemoglobin levels apparently normal, only with B12 levels extremely high. This suggests that this group of individuals is taking some king of supplements. The abnormally high levels of cobalamin seem to be the clustering criterion in this case.

Nevertheless, the filtered search for values under the reference was made. This cluster contains two patients with low folate levels, both with depression, one of them also has low hemoglobin.

Cluster 3

The third cluster has 22 individuals and just by observing the mean parameter values they appear to be healthy. The higher folate levels may be caused by supplementation.

A closer look at the data in Microsoft Excel reveals two cases of low cobalamin, one in a gastrointestinal and one in a cardiovascular patient; five cases of low iron in patients with gastrointestinal disease, diabetes and anemia and three cases of low hemoglobin with anemia, thyroid problems and diabetes.

Cluster 4

Finally the fourth cluster with 96 patients where 27 suffer from anemia and 11 from diabetes. The predominance of these illnesses perhaps justifies the below average mean levels of folate, B12, iron and hemoglobin.

As for the values below acceptance, low folate levels are registered in four cases of, diabetes, depression (neurological disease), autoimmune and rheumatic disease.

Low levels of cobalamin were seen in thirteen patients four with anemia, four with diabetes, two with allergies, two with thyroid complications and one with gastrointestinal disease.

Twenty eight persons have iron levels below the reference values, these are related to anemia, diabetes, autoimmune, cardiovascular disease, gastrointestinal disease, neurological disease, rheumatic disease, neoplasia and thyroid problems.

4.5.2. Ages from 60 to 79 years

This is the first elderly group, the dendogram corresponding to this group can be consulted in appendix B, figure B.2. For this group the cluster analysis returned four clusters, the results can be analyzed in tables 4.14 and 4.15.

Looking ate table 4.14 one should notice that the total hemoglobin level is below the reference value of 12 g/dL, which means most of the population has hemoglobin levels below the acceptable limits.

Cluster 1

Examining table 4.14 and 4.15 we can see the first cluster has more individuals than the other three. This cluster has folate levels below the average level and all other parameters have

normal mean levels. There is a predominance of patients with cardiovascular and gastrointestinal disease in this cluster and that might be caused by the folate levels.

Table 4.14 - Cluster analysis results for 60 - 79 years

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Folate (ng/mL)	Mean	8.45	7.68	11.82	21.13	11.02
	St Dev	2.73	2.91	5.47	3.83	6.03
B12 (pg/mL)	Mean	414.85	444.00	1724.80	526.97	525.81
	St Dev	168.26	210.97	266.18	262.65	375.78
Iron (µg/dL)	Mean	82.49	32.41	45.30	75.10	69.41
	St Dev	26.03	17.67	23.40	45.60	35.42
Hemoglobin (g/dL)	Mean	12.62	10.04	11.82	12.05	11.98
	St Dev	1.29	1.51	1.89	1.68	1.74

Table 4.15 - Number of persons with each illness aged 60 - 79 years

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Total	87	29	10	32	158
Anemia	4	2	3	9	18
Anti-Inflammatory		1			1
Diabetes	22	13	4	8	47
Cardiovascular Disease	36	8	2	11	57
Gastrointestinal Disease	9	2		2	13
Liver Disease				1	1
Neurological Disease	8	3			11
Rheumatic Disease	4		1	1	6
Thyroid	4				4

Once more, using the Microsoft Excel filter option, in this first cluster we find ten individuals with low serum folate (below 5 ng/mL), four were registered in cases of cardiovascular disease, three in diabetic patients, two in cases of anemia and one in a thyroid complication.

Low levels of vitamin B12 are observed in ten individuals, four cases of diabetes, three of cardiovascular disease, two of gastrointestinal disease and one case of anemia.

For iron, there are four concentrations below the reference values in table 3.1, they are associated with two cardiovascular disease, one diabetes and one gastrointestinal disease. There is also one rheumatic disease patient with iron levels over the limit.

Twenty seven cases of low hemoglobin levels were observed in this cluster, eleven in cardiovascular disease patients, eight in diabetics, two in anemia, neurological and rheumatic disease patients, one gastrointestinal patient and one suffering from thyroid related problems.

Cluster 2

In the second cluster, with 29 individuals, is characterized by a folate concentration below the population means and iron and hemoglobin below the reference limits. B12 levels are normal. This cluster has 13 individuals with diabetes and eight with cardiovascular disease as the most represented illness groups.

Analogously to the other clusters, low concentrations of folate were filtered so we can see what illnesses they relate to. In this cluster we found seven cases of low folate in three diabetics, one anemia patient, one gastrointestinal, one cardiovascular and one neurological disease.

Two cardiovascular patients, two diabetics and one individual taking anti-inflammatory medication form the low cobalamin group.

As for iron, below the reference concentrations were seen in sixteen individuals, six of those have diabetes, five have cardiovascular disease, two have anemia, one is taking anti-inflammatory medication, one has gastrointestinal and one has neurological disease.

Every individual in this cluster exhibits hemoglobin levels below the reference value.

Cluster 3

The third cluster has normal folate levels, as well as iron levels that are low but still above the reference. Hemoglobin is once more below the critical value and B12 is extremely high, again probably due to supplementation. This cluster was probably agglomerated with the high B12 as criterion.

The low levels analysis reveals low iron in two diabetes, one anemia and one rheumatic disease patient, and six cases of low hemoglobin with diabetes, anemia, disease and rheumatic disease.

Cluster 4

Cluster number four has normal concentrations for iron and cobalamin, hemoglobin just above the reference value and a higher than usual folate concentration. Analogously to the high B12 levels in the third cluster, one might suspect the persons in this cluster are taking folate supplements.

Using MS Excel to filter high and low concentrations, we find two cases of diabetes and two cases of cardiovascular disease associated with cobalamin levels under the reference values.

Low hemoglobin occurred in 14 patients and is associated with six cardiovascular patients, five anemic patients and three diabetics.

4.5.3. Ages 80 years and older

This last group also consists of elderly citizens, but these are in later stage of life where the body's functions are deteriorated, which in turn is compensated by medication. So for this group results are expected to be less clear, nonetheless conclusive results are expected.

For this group, there were found also four clusters. The dendogram can be consulted in appendix B, figure B.3, and the results in tables 4.16 and 4.17. The fact that there are 4 clusters for every age group is not premeditated, it was so dictated by the data.

Like the previous age group, persons over 80 years of age have total hemoglobin concentrations below the reference values.

Table 4.16 - Cluster analysis results for >80 years

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Folate (ng/mL)	Mean	10.07	7.13	9.78	21.78	12.27
	St Dev	4.31	1.85	6.25	5.75	6.77
B12 (pg/mL)	Mean	377.15	444.86	2001.00	719.94	543.20
	St Dev	121.19	107.56	0.00	363.33	421.25
Iron (µg/dL)	Mean	61.44	105.71	39.00	93.81	71.28
	St Dev	23.32	23.43	12.25	38.74	32.27
Hemoglobin (g/dL)	Mean	11.68	15.34	11.25	11.26	11.91
	St Dev	1.47	1.03	1.94	1.34	1.80

Table 4.17 - Number of persons with each illness aged >80 years

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Total	48	7	4	16	75
Anemia	3			3	6
Diabetes	12	1	2	4	19
Cardiovascular Disease	22	3	2	1	28
Gastrointestinal Disease	5	1		1	7
Neurological Disease	3	2		1	6
Renal Disease	1			2	3
Respiratory Disease				1	1
Rheumatic Disease				1	1
Neoplasia	1			1	2
Thyroid	1			1	2

Cluster 1

Analysis of tables 4.16 and 4.17 reveal a first cluster with forty eight patients, twenty two with cardiovascular disease and twelve with diabetes are the predominant groups. As for the concentrations, they all seem normal except hemoglobin, which is below the reference levels.

A closer look into this cluster, using the filtering tool in MS Excel, reveals five folate concentrations below 5 ng/mL, registered in two patients with diabetes, two with cardiovascular and one with renal disease.

Cobalamin levels under the reference values in table 3.1 occur in nine patients, 5 with cardiovascular disease, two with diabetes, one with anemia and another with neurological disease.

Iron registers low levels in nine persons, four suffering from cardiovascular disease, two from diabetes, one with anemia, one from gastrointestinal and another from neurological disease.

Hemoglobin levels are below reference values in twenty nine patients, fourteen of those register cardiovascular disease, seven diabetes, four gastrointestinal disease, two anemia and another two neurological disease.

Cluster 2

The second cluster has folate levels lower than the total average, high iron concentrations but not above the reference values, normal B12 and hemoglobin. This is a small cluster of only

seven individuals. Where on a closer examination every individual has healthy values except for one with coeliac disease with serum folate under 5 ng/mL.

Cluster 3

Another small cluster, number three, has only four individuals, two diabetics and two cardiovascular disease patients. Folate and iron levels are below average and hemoglobin is below acceptable values. Cobalamin concentrations are extremely high, once more suggesting supplementation with vitamin B12.

Filtering for under the reference values in table 3.1, one of the diabetic patients has low concentrations of folate and one diabetic and one cardiovascular patients have low hemoglobin.

Cluster 4

Cluster four shows levels of folate higher than the population means, which may be caused by the ingestion of supplements. B12 and iron show normal levels but hemoglobin concentrations are low.

Using MS Excel's filter, we find there is one diabetic with high iron concentration, over the reference value in table 3.1. Low levels of hemoglobin occur in almost every patient except the one with cardiovascular disease and the one with gastrointestinal disease (anorexia).

4.6. Sensitivity Analysis

This sensitivity analysis will help assess the power of the cluster analysis performed earlier. As it was demonstrated in table 4.11, groups 3 and 4, corresponding respectively to groups with 60 to 79 years and 80 years and above (elderly groups), are not statistically significantly different. So, for this last part of the study a cluster analysis was performed merging together groups 3 and 4, so the ages in this group are 60 years and over.

The dendogram for this cluster analysis can be viewed in appendix C. Six clusters were formed and the results can be consulted in tables 4.18 and 4.19.

As we can see in tables 4.18, the previous tendency for below reference values of hemoglobin is maintained, which corroborates the separated cluster analysis conclusion that hemoglobin levels are very low in elderly citizens.

Cluster 1

In the first cluster it is possible to see that folate levels are below the populations mean value. Hemoglobin is below the reference value (12 g/dL). Vitamin B12 levels are slightly lower

but considerably over the reference value of 273 pg/mL. Iron levels are normal. This cluster has 62 persons, where we highlight 16 diabetes cases and 29 cardiovascular diseases.

Table 4.18 - Sensitivity cluster analysis results for >60 years

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Total
Folate (ng/mL)	Mean	7.98	7.84	9.85	11.24	19.78	24.11	11.43
	St Dev	2.85	2.98	3.75	5.53	4.32	2.20	6.29
B12 (pg/mL)	Mean	359.79	456.16	450.83	1803.71	468.44	699.53	531.43
	St Dev	105.23	193.40	193.54	256.55	191.21	396.90	390.23
Iron (µg/dL)	Mean	65.50	34.66	95.07	43.50	48.74	110.32	70.02
	St Dev	18.68	17.85	22.15	20.56	23.05	46.48	34.38
Hemoglobin (g/dL)	Mean	11.79	9.97	13.42	11.66	11.37	11.96	11.96
	St Dev	0.80	1.50	1.25	1.85	1.67	1.54	1.75

Table 4.19 - Number of persons with each illness aged >60 years

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Total
Total	62	38	72	14	27	19	232
Anemia	3	2	4	3	7	5	24
Anti-inflammatory		1					1
Diabetes	16	17	15	6	6	6	66
Cardiovascular Disease	29	12	29	4	7	4	85
Gastrointestinal Disease	7	2	7		2	2	20
Liver Disease			1				1
Neurological Disease	4	4	7		2		17
Renal Disease			1		2		3
Respiratory Disease					1		1
Rheumatic Disease	1		3	1		1	6
Neoplasia			1			1	2
Thyroid	2		4				6

Cluster 2

The second cluster, like the first one, shows folate levels below the population average. This time hemoglobin is extremely low. Iron levels are also below reference limits (37 – 145 µg/dL). Cobalamin concentrations are normal. This cluster has dangerously low values for most of the parameters and is associated with diabetes, cardiovascular disease, conditions most represented in this cluster.

Cluster 3

Cluster number three shows folate concentrations below the means, although not as low as clusters one and two. Mean levels for the remaining parameters are normal, with emphasis on hemoglobin, it's the only cluster whit concentrations above the acceptable values. Like the previous clusters the most common illnesses are diabetes, cardiovascular disease, gastrointestinal disease and neurological disease.

Cluster 4

The low hemoglobin level trend returns in the fourth cluster. In this cluster the folate levels are normal and iron concentrations are over the reference value, still slightly under the average for this population. Vitamin B12 levels stand out as they are extremely high, suggesting that this group may be under supplementation.

Cluster 5

Cluster five exhibits hemoglobin and iron concentrations similar to those in cluster four. Cobalamin levels are normal. Folate levels are high, which is good, but a careful look at some of the values arises the suspicion of supplementation. Anemia, diabetes and cardiovascular disease are the main illnesses. Patients suffering from anemia are often supplemented with folate.

Cluster 6

The sixth and last cluster has B12 and iron levels above the average for this population, but within the normal values. Hemoglobin mean continues to be below the reference value. Similarly to the last cluster folate levels are high, and again this is probably due to supplementation since most of the values are over the measuring range (>25 ng/mL).

Since the data is the same, the associations between low or high biomarker concentrations and the illnesses, that were assessed using Microsoft Excel filters, are the same as the ones described in the cluster analysis presented earlier.

5. Conclusions and Recommendations

Multivariate analysis proved to be a useful tool to characterize a population based on clinical analysis data. Helping find correlations and groups in data that were otherwise unassessable.

5.1. General conclusions

The descriptive statistics revealed that the population's folate levels were low. Most of the values were below the average, and 13.5% of the population has values below the 5 ng/ml value. Concentrations of this biomarker increase with age.

Vitamin B12 and Iron mean levels fall between the reference range (>273 pg/mL for B12 and 37 – 145 µg/dL for iron). However 13.3% and 18% of the population register respectively cobalamin and iron values below the reference.

Hemoglobin is dangerously low in this population, with 38.9% of the population registering levels below the reference value of 12 g/dL.

With the spearman correlation we were able to find correlations between iron, ferritin and transferrin, and iron with hemoglobin. Transferrin is inversely correlated with iron, which supports the fact that it decreases in high concentrations of iron.

Kruskal-Wallis analysis of variance demonstrated that the defined age groups were significantly different from each other, except from the two elderly groups (60 to 79 years and over 80 years). This demonstrates that the biomarkers' patterns alter with age.

The second age group, containing persons from 30 to 59 years of age, has the most number of low values for every biomarker.

5.2. Cluster analysis conclusions

The first conclusion we can take from the cluster analysis is that although most folate values are above the reference, most of them are below the average for the population. Given that some of the individuals in the population are being supplemented, and this boosts the means for folate, this reveals that in general folate levels are low for this population.

Another fact that shows clearly in the cluster analysis is that hemoglobin is in general very low for this population. Unlike folate, these levels are not just low, most of them are under the reference values and represent a hazard for the health of these individuals. This characteristic is most critic in elderly patients that reveal an overall hemoglobin mean below the reference value.

Iron and vitamin B12 levels for this population are mostly normal and within the reference ranges. There are some cases of low concentrations of both parameters, but most individuals are healthy in regard to cobalamin and iron.

The influence of the biomarkers in certain health complications described in the literature review, has been corroborated by the cluster analysis as low levels of folate, cobalamin, iron or hemoglobin have shown connections with the diseases, mainly cardiovascular, gastrointestinal, neurological, diabetes and anemia.

5.3. Sensitivity analysis conclusions

The sensitivity analysis provided a second look at the variables from a different point of view. The results are similar to the ones observed in the cluster analysis. Clusters were formed according to the parameter concentrations, and even with more data the clustering tendency is the same. It is still possible to identify the groups with individuals with low parameter concentrations and the ones that we suspect are supplemented.

Folate levels are still low even though they do not reach dangerous levels. Hemoglobin still shows very low levels, this becomes clear when in 232 individuals only a cluster of 72 has a positive mean value. Analogously to the first cluster analysis, vitamin B12 and iron exhibit normal values in most cases.

Both in the cluster analysis and sensitivity analysis, no evidence was found that the parameters are interdependent, for example low folate levels do not implicate low values for

any of the other parameters, and vice versa. This corroborates the results from the Spearman correlation analysis.

5.4. Recommendations

The first thing that I would recommend for a future study, is to use a control group. It is important to have an unbiased reference for comparison of values, and a control group can help define what is "normal" for a certain population. Such a group could also help with the blur around the reference values by helping the researcher to define his own reference for his population.

Another fact that would help with this type of study is to obtain more information about the patients, mainly eating and lifestyle habits, and supplement intake. It is proven that food intake, smoking or drinking can affect the folate, B12, iron or hemoglobin concentrations. Taking these factors into account would produce a more accurate study.

Bibliography

Abonyi, J. & Feil, B., 2007. *Cluster Analysis for Data Mining and System Identification*, Springer. Available at: http://books.google.com/books?id=SUMoMgPDY2AC&pgis=1 [Accessed November 28, 2013].

Almeida, C. et al., 2012. Serum folic acid is reduced in patients with Alzheimer's disease Redução dos níveis séricos de ácido fólico em pacientes com a doença de Alzheimer. *Revista de Psiquiatria Clínica*, (Mci), pp.18–21. Available at: http://www.hcnet.usp.br/ipq/revista/vol39/n3/eng/90.htm [Accessed January 16, 2014].

Andrès, E. et al., 2007. Clinical aspects of cobalamin deficiency in elderly patients. Epidemiology, causes, clinical manifestations, and treatment with special focus on oral cobalamin therapy. *European journal of internal medicine*, 18(6), pp.456–62.

Arabie, P., Hubert, L.J. & De Soete, G., 1996. *Clustering and Classification*, World Scientific. Available at: http://books.google.com/books?id=AgovP9eJPsUC&pgis=1 [Accessed November 28, 2013].

Baik, H., 1999. Vitamin B12 deficiency in the elderly. *Annual review of nutrition*, 19, pp.357–377.

Brolin, R.E. et al., 1998. Are vitamin B12 and folate deficiency clinically important after roux-en-Y gastric bypass? *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract*, 2(5), pp.436–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9843603.

Burtis, C.A., Ashwood, E.R. & Bruns, D.E., 2012. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics*, Available at: http://www.google.pt/books?hl=pt-PT&lr=&id=BBLRUI4aHhkC&pgis=1 [Accessed November 25, 2013].

Cabrera, C. et al., 2007. Socio-economic gradient in food selection and diet quality among 70-year olds. *The journal of nutrition, health & aging*, 11(6), pp.466–73.

Campbell, A.K. et al., 2003. Plasma Vitamin B-12 Concentrations in an Elderly Latino Population Are Predicted by Serum Gastrin Concentrations and Crystalline Vitamin B-12 Intake. *Journal of nutrition*, (February), pp.2770–2776.

Carrilho, M. & Patrício, L., 2004. A situação demográfica recente em Portugal. *Revista de Estudos Demográficos*, pp.101–146.

Christensen, B. et al., 1999. Whole blood folate, homocysteine in serum, and risk of first acute myocardial infarction. *Atherosclerosis*, 147(2), pp.317–26.

Chui, C.H. et al., 2001. Vitamin B12 deficiency--need for a new guideline. *Nutrition (Burbank, Los Angeles County, Calif.)*, 17(11-12), pp.917–20.

Clarke, R. et al., 1998. Folate, vitamin B12, and serum total homocysteine levels in confirmed Alzheimer disease. *Archives of neurology*, 55(11), pp.1449–55. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9823829.

Clarke, R. et al., 2003. Screening for vitamin B-12 and folate deficiency in older persons. *The American journal of clinical nutrition*, 77, pp.1241–1247.

Clarke, R. et al., 2004. Vitamin B12 and Folate Deficiencіy in Later Life. *Care of Gastrointestinal Problems in the Older Adult. …*, 33(1), pp.34–41. Available at: http://www.ageing.oxfordjournals.org/cgi/doi/10.1093/ageing/afg109 [Accessed January 21, 2014].

Coladarci, T. et al., 2010. *Fundamentals of Statistical Reasoning in Education*, John Wiley & Sons. Available at: http://books.google.com/books?id=due_H9uyOaUC&pgis=1 [Accessed February 26, 2014].

Coussement, K., Demoulin, N. & Charry, K., 2011. *Marketing Research with SAS Enterprise Guide*, Gower Publishing, Ltd. Available at: http://books.google.com/books?id=6T9pkCbt_WEC&pgis=1 [Accessed November 25, 2013].

Czeizel, A.E. & Vargha, P., 2004. Periconceptional folic acid/multivitamin supplementation and twin pregnancy. *American journal of obstetrics and gynecology*, 191(3), pp.790–4.

Deep, R., 2006. *Probability and Statistics with Integrated Software Routines*, Academic Press. Available at: http://books.google.com/books?id=FCax7O6U6NQC&pgis=1 [Accessed February 25, 2014].

Everitt, B.S. et al., 2011. *Cluster Analysis (Google eBook)*, John Wiley & Sons. Available at: http://books.google.com/books?id=w3bE1kqd-48C&pgis=1 [Accessed November 28, 2013].

Ferraty, F. & Vieu, P., 2006. *Nonparametric Functional Data Analysis: Theory and Practice (Google eBook)*, Springer. Available at: http://books.google.com/books?id=lMy6WPFZYFcC&pgis=1 [Accessed February 24, 2014].

Gibson, R.S., 2005. *Principles of Nutritional Assessment*, Oxford University Press. Available at: http://books.google.com/books?id=lBlu7UKI3aQC&pgis=1 [Accessed January 27, 2014].

Gravetter, F. & Wallnau, L., 2013. *Essentials of Statistics for the Behavioral Sciences*, Cengage Learning. Available at: http://books.google.com/books?id=LC9xc3CqpwIC&pgis=1 [Accessed February 25, 2014].

Hair, J.F. et al., 2007. *Analise Multivariada de Dados*, Bookman. Available at: http://books.google.com/books?id=LxFb5JzXdbUC&pgis=1 [Accessed November 19, 2013].

Herbert, V., 1987. Making sense of laboratory tests of folate status: folate requirements to sustain normality. *American journal of hematology*, (26), pp.199–207. Available at: http://onlinelibrary.wiley.com/doi/10.1002/ajh.2830260211/abstract [Accessed February 7, 2014].

Hilton, J.J., 2007. A comparison of folic acid awareness and intake among young women aged 18-24 years. *Journal of the American Academy of Nurse Practitioners*, 19(10), pp.516–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17897115 [Accessed February 18, 2014].

Instituto Nacional de Estatística, I., 2011. *Estatísticas Demográficas 2011*,

Iyer, R. & Tomar, S.K., 2009. Folate: a functional food constituent. *Journal of food science*, 74(9), pp.R114–22.

Johnson, R. & Wichern, D., 2007. *Applied multivariate statistical analysis*, Available at: http://cisco.qu.edu.qa/artssciences/mathphysta/stats/syllabi/Syllabus- spring 2012/Statistics/Dr_Alodat_STAT_459_L01_Spring_2012.pdf [Accessed December 10, 2013].

Kaufman, L. & Rousseeuw, P.J., 2009. *Finding Groups in Data: An Introduction to Cluster Analysis (Google eBook)*, John Wiley & Sons. Available at: http://books.google.com/books?id=YeFQHiikNo0C&pgis=1 [Accessed November 28, 2013].

Kerr, M. a et al., 2009. Folate, related B vitamins, and homocysteine in childhood and adolescence: potential implications for disease risk in later life. *Pediatrics*, 123(2), pp.627–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19171631 [Accessed February 5, 2014].

King, B.M., Rosopa, P.J. & Minium, E.W., 2011. *Statistical Reasoning in the Behavioral Sciences*, John Wiley and Sons. Available at:

http://books.google.com/books?id=9_imyHBaZyMC&pgis=1 [Accessed March 6, 2014].

Kruskal, W.H. & Wallis, W.A., 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical …*, 47(260), pp.583–621. Available at: http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441 [Accessed March 6, 2014].

Kvam, P.H. & Vidakovic, B., 2007. *Nonparametric Statistics with Applications to Science and Engineering (Google eBook)*, John Wiley & Sons. Available at: http://books.google.com/books?id=K7maw0lvu7kC&pgis=1 [Accessed February 25, 2014].

Lancellotti, C.C., 2012. Revisión sistemática del efecto de los folatos y otros nutrientes relacionados en la función cognitiva del adulto mayor. *Nutrición Hospitalaria*, 27(1), pp.90–102.

Liappas, I. a et al., 2007. Vitamin B12 and hepatic enzyme serum levels correlate with interleukin-6 in alcohol-dependent individuals without liver disease. *Clinical biochemistry*, 40(11), pp.781–6.

Lima-Costa, M.F., 2012. Nível de hemoglobina entre idosos e sua associação com indicadores do estado nutricional e uso de serviços de saúde : Projeto Bambuí Hemoglobin level in older adults and the association with nutritional status and use of health services : the Bambuí Proje. , 28(11), pp.2085–2094.

Malhotra, N.K., 2006. *Pesquisa de Marketing: Uma Orientação*, Bookman. Available at: http://books.google.com/books?id=FtdIFOgTP8UC&pgis=1 [Accessed November 22, 2013].

Matthews, J.H., 1995. Cobalamin and folate deficiency in the elderly. *Baillière's clinical haematology*, 8(3), pp.679–97.

Mazeh, D. et al., 2005. Elderly psychiatric patients at risk of folic acid deficiency: a case controlled study. *Archives of gerontology and geriatrics*, 41(3), pp.297–302.

Mccully, K., 1999. Homocysteine, vitamin deficiency and prevention of arteriosclerosis. *Integrative Medicine*, 1(1), pp.3–9.

Mesquita, M., 2009. *Estimativa da estrutura em portugueses com idade igual ou superior a 60 anos*. Master Thesis. Faculdade de Medicina da Universidade de Coimbra, Coimbra.

Mitrushina, M. et al., 2005. *Handbook of Normative Data for Neuropsychological Assessment*, Oxford University Press. Available at: http://books.google.com/books?id=Ygndi8UlmxkC&pgis=1 [Accessed February 26, 2014].

Montgomery, D.C. & Runger, G.C., 2010. *Applied Statistics and Probability for Engineers*, John Wiley & Sons. Available at: http://books.google.com/books?id=_f4KrEcNAfEC&pgis=1 [Accessed November 12, 2013].

De Muth, J.E., 2006. *Basic Statistics and Pharmaceutical Statistical Applications, Second Edition*, CRC Press. Available at: http://books.google.com/books?id=Lj5PU_psLCMC&pgis=1 [Accessed February 26, 2014].

O'Rourke, N., Hatcher, L. & Stepanski, E.J., 2005. *A Step-by-step Approach to Using SAS for Univariate & Multivariate Statistics*, SAS Institute. Available at: http://books.google.com/books?id=pfUfzykTZ1AC&pgis=1 [Accessed February 25, 2014].

Okumura, K. & Tsukamoto, H., 2011. Folate in smokers. *Clinica chimica acta; international journal of clinical chemistry*, 412(7-8), pp.521–6.

Osborn, C.E., 2000. *Statistical Applications for Health Information Management*, Jones & Bartlett Learning. Available at: http://books.google.com/books?id=TmgFOJSUukwC&pgis=1 [Accessed March 6, 2014].

Pepper, M.R. & Black, M.M., 2011. B12 in fetal development. *Seminars in cell & developmental biology*, 22(6), pp.619–23.

Pereira, J.C.R., 1999. *Análise de Dados Qualitativos: Estratégias Metodológicas para as Ciências da Saúde Humanas e Sociais*, EdUSP. Available at: http://books.google.com/books?id=uoBSa0IsA8QC&pgis=1 [Accessed November 28, 2013].

Pett, M.A., 1997. *Nonparametric Statistics in Health Care Research: Statistics for Small Samples and Unusual Distributions*, SAGE. Available at: http://books.google.com/books?id=yU15rUiLRI8C&pgis=1 [Accessed January 22, 2014].

Plichta, S.B. & Garzon, L.S., 2009. *Statistics for Nursing and Allied Health*, Lippincott Williams & Wilkins. Available at: http://books.google.com/books?id=yc-WIr2dZBwC&pgis=1 [Accessed March 6, 2014].

Ramos, M.I. et al., 2005. Low folate status is associated with impaired cognitive function and dementia in the Sacramento Area Latino Study on Aging. *The American journal of clinical nutrition*, 82(6), pp.1346–52.

Refsum, H. et al., 2004. Facts and recommendations about total homocysteine determinations: an expert opinion. *Clinical chemistry*, 50(1), pp.3–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14709635 [Accessed January 13, 2014].

Reynolds, E., 2006. Vitamin B12, folic acid, and the nervous system. *Lancet neurology*, 5(11), pp.949–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17052662.

Roddie, C. & Davis, B., 2009. Iron, B12 and folate. *Medicine*, 37(3), pp.125–128.

Romesburg, C., 2004. *Cluster Analysis for Researchers*, Lulu.com. Available at: http://books.google.com/books?id=ZuIPv7OKm10C&pgis=1 [Accessed November 28, 2013].

Rubin, A., 2012. *Statistics for Evidence-Based Practice and Evaluation*, Cengage Learning. Available at: http://books.google.com/books?id=UQoKAAAAQBAJ&pgis=1 [Accessed February 26, 2014].

Semmelrock, M.J. et al., 2012. Reticulocyte hemoglobin content allows early and reliable detection of functional iron deficiency in blood donors. *Clinica chimica acta; international journal of clinical chemistry*, 413(7-8), pp.678–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22212625 [Accessed February 27, 2014].

Serviço de Estudos sobre a População Departamento de Estatísticas Censitárias e da População, S., 2002. *O Envelhecimento em Portugal : Situação demográfica e socio-económica recente das pessoas idosas*,

Sheskin, D.J., 2003. *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition (Google eBook)*, CRC Press. Available at: http://books.google.com/books?id=bmwhcJqq01cC&pgis=1 [Accessed January 22, 2014].

Shuaibi, A.M., House, J.D. & Sevenhuysen, G.P., 2008. Folate status of young Canadian women after folic acid fortification of grain products. *Journal of the American Dietetic Association*, 108(12), pp.2090–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19027414 [Accessed February 17, 2014].

Tamhane, A.C., 2009. *Statistical Analysis of Designed Experiments: Theory and Applications*, John Wiley & Sons. Available at: http://books.google.com/books?id=5fHm_OWUQmQC&pgis=1 [Accessed February 26, 2014].

Taneja, S. et al., 2007. Cobalamin and folate status in infants and young children in a low-to-middle income community in India. *The American journal of clinical nutrition*, 86(5), pp.1302–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17991639.

Viñas, B.R. et al., 2011. Projected Prevalence of Inadequate Nutrient Intakes in Europe. *Annals of Nutrition & Metabolism*, 59(2-4), pp.84–95.

Vining, G.G. & Kowalski, S.M., 2010. *Statistical Methods for Engineers*, Cengage Learning. Available at: http://books.google.com/books?id=htmWSYtornYC&pgis=1 [Accessed November 22, 2013].

Ward, J.H.J., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*. Available at: http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845 [Accessed December 9, 2013].

Weimer, J., 1998. Factors affecting nutrient intake of the elderly U.S. Department of Agriculture, ed. *U.S. Department of Agriculture - Agricultural Economics Reports*, (769), pp.1–10.

Weinberg, S.L. & Abramowitz, S.K., 2002. *Data Analysis for the Behavioral Sciences Using SPSS*, Cambridge University Press. Available at: http://books.google.com/books?id=lVR7YrycKbcC&pgis=1 [Accessed February 25, 2014].

Westfall, P.H. et al., 1999. *Multiple Comparisons and Multiple Tests: Using the SAS System*, SAS Institute. Available at: http://books.google.com/books?id=Ta5UzMajyu0C&pgis=1 [Accessed February 26, 2014].

Appendices

A. One-Way ANOVA Appendix

Table A.1 - One-way anova for vitamin B12

	SS	d.f.	MS	F	*p*-value
Intercept	72781362	1	72781362	642.94	0.000
Age	534665	3	178222	1.57	0.195
Error	46865319	414	113201		

Table A.2 - Tests for ANOVA variance assumptions for vitamin B12

	Hartley F-max	Cochran C	Bartlett Chi-sqr	d.f.	*p*-value
B12	3.565	0.408	39.140	3	0.000

Table A.3 - Kruskal-Wallis medians test for vitamin B12

	Age Group	1	2	3	4	Total
Median	Observed	20.00	74.00	79.00	36.00	209.00
	Expected	17.00	76.00	78.50	37.50	
	Obs.-Exp.	3.00	-2.00	0.50	-1.50	
Median	Observed	14.00	78.00	78.00	39.00	209.00
	Expected	17.00	76.00	78.50	37.50	
	Obs.-Exp.	-3.00	2.00	-0.50	1.50	
Total	Observed	34.00	152.00	157.00	75.00	418.00

Table A.4 - Kruskal_Wallis significantly different groups for vitamin B12 (*p*-values)

Age Group	1	2	3	4
1		1.000	1.000	1.000
2	1.000		1.000	1.000
3	1.000	1.000		1.000
4	1.000	1.000	1.000	

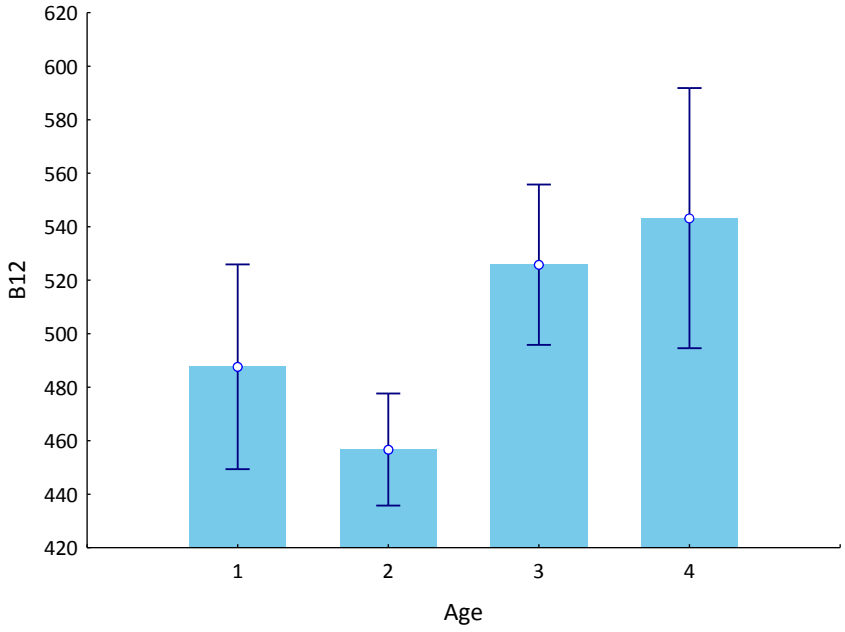

Figure A.1 - Vitamin B12 levels for each age group (the bars represent the means and the whiskers the standard deviations)

Table A.5 - One-way ANOVA for iron

	SS	d.f.	MS	F	*p*-value
Intercept	1579744	1	1579744	934.00	0.000
Age	8155	3	2718	1.61	0.187
Error	700232	414	1691		

Table A.6 - Tests for ANOVA variance assumptions for iron

	Hartley F-max	Cochran C	Bartlett Chi-sqr	d.f.	*p*-value
B12	2.602	0.374	24.410	3	0.000

Table A.7 - Kruskal-Wallis medians test for iron

	Age Group	1	2	3	4	Total
Median	Observed	18.00	74.00	79.00	39.00	210.00
	Expected	17.08	76.36	78.88	37.68	
	Obs.-Exp.	0.92	-2.36	0.12	1.32	
Median	Observed	16.00	78.00	78.00	36.00	208.00
	Expected	16.92	75.64	78.12	37.32	
	Obs.-Exp.	-0.92	2.36	-0.12	-1.32	
Total	Observed	34.00	152.00	157.00	75.00	418.00

Table A.8 - Kruskal-Wallis significantly different groups for iron (*p*-values)

Age Group	1	2	3	4
1		1.000	1.000	1.000
2	1.000		0.860	1.000
3	1.000	0.860		1.000
4	1.000	1.000	1.000	

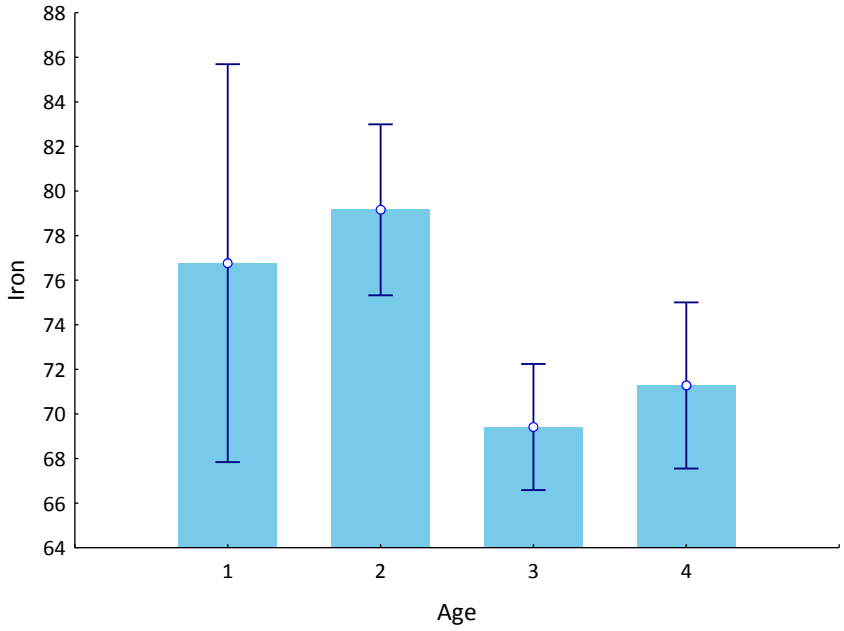

Figure A.2 - Iron levels for each age group (the bars represent the means and the whiskers the standard deviations)

Table A.9 - One-way ANOVA for hemoglobin

	SS	d.f.	MS	F	*p*-value
Intercept	42698.63	1	42698.63	14338.88	0.000
Age	22.49	3	7.50	2.52	0.058
Error	1232.82	414	2.98		

Table A.10 - Tests for ANOVA variance assumptions

	Hartley F-max	Cochran C	Bartlett Chi-sqr	d.f.	*p*-value
B12	1.165	0.266	0.666	3	0.881

Table A.11 - Kruskal-Wallis medians test for hemoglobin

	Age Group	1	2	3	4	Total
Median	Observed	15.00	69.00	90.00	50.00	224.00
	Expected	18.22	81.45	84.13	40.19	
	Obs.-Exp.	-3.22	-12.45	5.87	9.81	
Median	Observed	19.00	83.00	67.00	25.00	194.00
	Expected	15.78	70.54	72.87	34.81	
	Obs.-Exp.	3.22	12.45	-5.87	-9.81	
Total	Observed	34.00	152.00	157.00	75.00	418.00

Table A.12 - Kruskal-Wallis significantly different groups for hemoglobin (*p*-values)

Age Group	1	2	3	4
1		1.000	1.000	0.383
2	1.000		0.151	**0.041**
3	1.000	0.151		1.000
4	0.383	**0.041**	1.000	

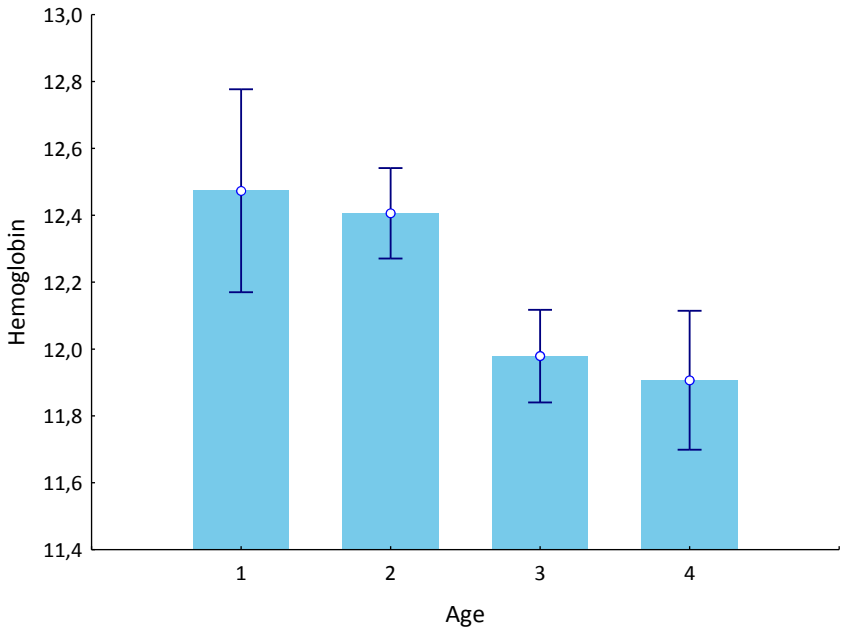

Figure A.3 - Hemoglobin levels for each age group (the bars represent the means and the whiskers the standard deviations)

B. Cluster Analysis Appendix

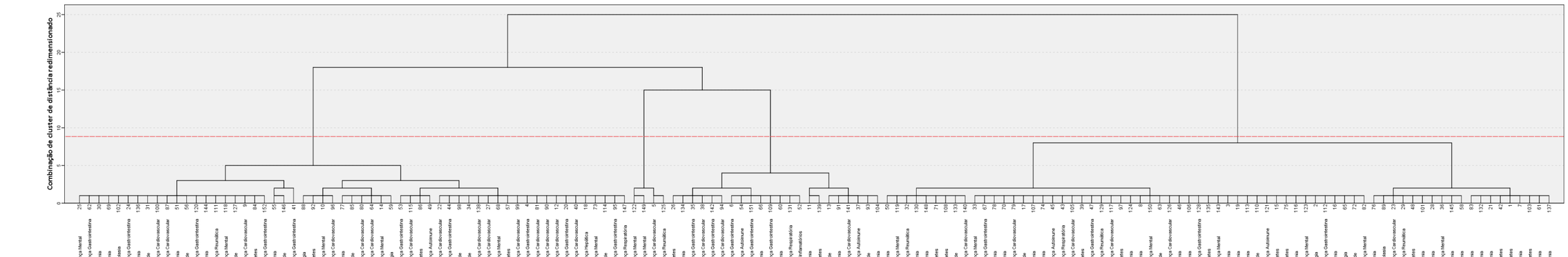

Figure B.1 - Dendogram for ages 30 to 59 years

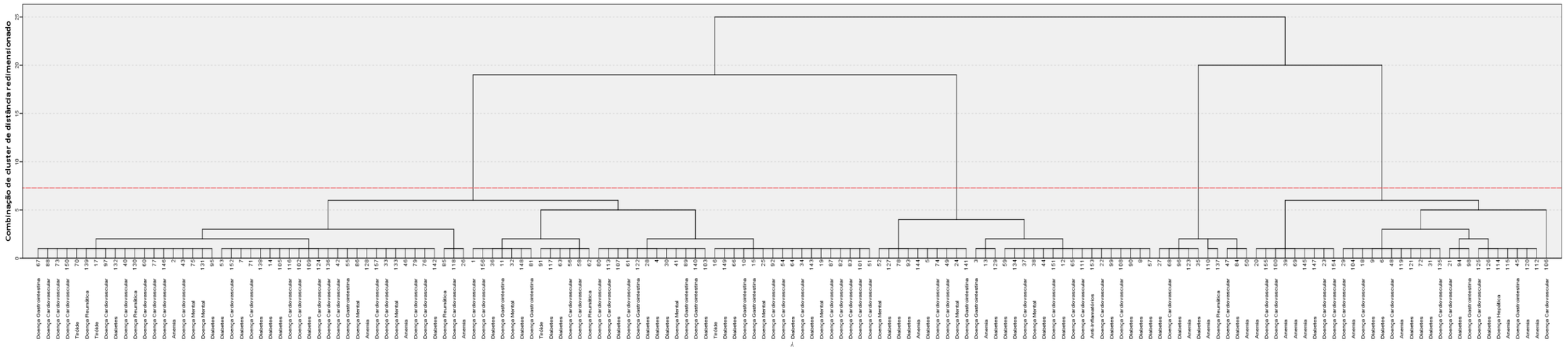

Figure B.2 - Dendogram for ages 60 to 79 years

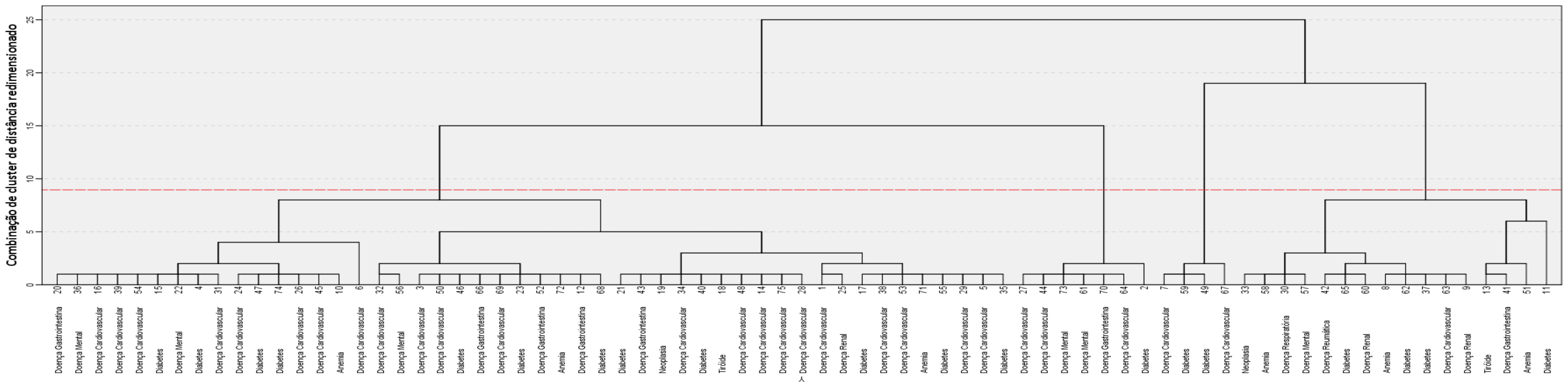

Figure B.3 - Dendogram for ages 80 years and over

C. *Sensitivity Analysis Appendix*

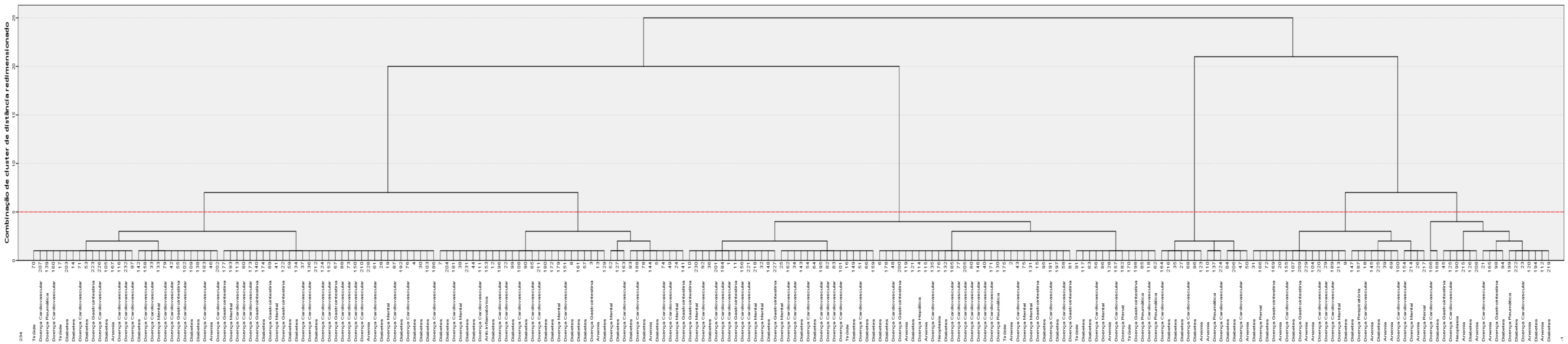

Figure C.1 - Dendogram for ages 60 years and over