NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

INTERDISCIPLINARY POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MSc THESIS

Data Analytics in Chronic Disease Self-Management

Statistical and Machine Learning Methodologies
for Knowledge Discovery based on Quantified Self Data

Aikaterini V. Georgountzou

Supervisor: **Dr. Anastasia Krithara,** Post-Doctoral Research Associate, Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, NCSR “Demokritos”

ATHENS

FEBRUARY 2019

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Η Ανάλυση Δεδομένων στην Αυτοδιαχείριση Χρόνιων Παθήσεων

Μεθοδολογίες Στατιστικής και Μηχανικής Μάθησης για Εξόρυξη Γνώσεων
από Προσωπικά Δεδομένα Υγείας και Τρόπου Ζωής

Αικατερίνη Β. Γεωργούντζου

Επιβλέπουσα: **Δρ. Αναστασία Κριθαρά,** Μεταδιδακτορική Ερευνήτρια,
Εργαστήριο Τεχνολογίας Γνώσεων και Λογισμικού, Ινστιτούτο
Πληροφορικής και Τηλεπικοινωνιών, ΕΚΕΦΕ «Δημόκριτος»

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2019

MSc THESIS

Data Analytics in Chronic Disease Self-Management: Statistical and Machine Learning Methodologies for Knowledge Discovery based on Quantified Self Data

Aikaterini V. Georgountzou

S.N.: ΠΙΒ0160

SUPERVISOR: **Dr. Anastasia Krithara,** Post-Doctoral Research Associate, Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, NCSR “Demokritos”

EXAMINATION COMMITTEE:

Dr. Anastasia Krithara, Post-Doctoral Research Associate, Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, NCSR “Demokritos”

Dr. Ilias Maglogiannis, Associate Professor, Director of the Computational Biomedicine Laboratory, Department of Digital Systems, University of Piraeus

Dr. Vangelis Karkaletsis, Research Director of the Institute of Informatics and Telecommunications, Head of the Software and Knowledge Engineering Laboratory, NCSR "Demokritos"

February 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Η Ανάλυση Δεδομένων στην Αυτοδιαχείριση Χρόνιων Παθήσεων: Μεθοδολογίες Στατιστικής και Μηχανικής Μάθησης για Εξόρυξη Γνώσεων από Προσωπικά Δεδομένα Υγείας και Τρόπου Ζωής

Αικατερίνη Β. Γεωργούντζου
Α.Μ.: ΠΙΒ0160

ΕΠΙΒΛΕΠΟΥΣΑ: **Δρ. Αναστασία Κριθαρά,** Μεταδιδακτορική Ερευνήτρια, Εργαστήριο Τεχνολογίας Γνώσεων και Λογισμικού, Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών, ΕΚΕΦΕ «Δημόκριτος»

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Δρ. Αναστασία Κριθαρά, Μεταδιδακτορική Ερευνήτρια, Εργαστήριο Τεχνολογίας Γνώσεων και Λογισμικού, Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών, ΕΚΕΦΕ «Δημόκριτος»

Δρ. Ηλίας Μαγκλογιάννης, Αναπληρωτής Καθηγητής, Διευθυντής Εργαστηρίου Υπολογιστικής Βιοϊατρικής, Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς

Δρ. Βαγγέλης Καρκλέτσης, Διευθυντής Έρευνας, Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών, Υπεύθυνος Εργαστηρίου Τεχνολογίας Γνώσεων και Λογισμικού, ΕΚΕΦΕ «Δημόκριτος»

Φεβρουάριος 2019

ABSTRACT

Chronic diseases management is one of the greatest challenges of modern healthcare systems. Given the fact that non-communicable diseases are responsible for more than 70% of deaths worldwide [1], the constant monitoring of a patient's health condition has become vital need and, hence, the era of mobile health starts to rise. At the same time, the idea of self-managing personal aspects of life, and not only, through the prism of new technologies, the so-called Quantified Self, gains ground rapidly. Nowadays, sensors constitute an integral part of the daily life and monitor almost every aspect of it, gathering enormous quantities of data. The challenge is how to control the data that derive from the combination of electronic health services with wearable sensor technologies and broaden the horizons of scientific research [2]. At this point, data analytics assumes its decisive role. Patients using such technologies gain the capability to record and process their vital signs, fitness activities, everyday habits, or even feelings [3]. The resulting data constitute the gemstone for statistical and machine learning techniques to be performed so that knowledge discovery can take place and, as a consequence, identify the risk factors in patients' health and provide personalized medical follow-up and immediate feedback to avoid emergent situations.

This graduate thesis proposes a data analytics solution that will examine patients' consistency in their measurement schedule and study the interaction among the different daily measurements, with the scope of determining how these factors can influence the monitoring of their health. Studies generalized on a demographic level, including sex, age and geolocation, will also take place so that statistical significant differences can be identified in the medical values and, thus, appropriate recommendations can be derived per population group. Aiming at improving and personalizing the medical monitoring of chronic health conditions, the proposed solution can circumvent the challenges of electronic health systems and provide benefits for the involved patients, such as enhancement of their welfare, early detection of dangerous situations, assumption of further targeted monitoring, motivation to engage in self-caring activities and follow treatment and, last, modeling of their behavior to improve self-care and enjoy a better quality of life.

SUBJECT AREA: Data Analytics

KEYWORDS: chronic diseases management, quantified self, mobile health, statistical data analysis, machine learning

ΠΕΡΙΛΗΨΗ

Η διαχείριση των χρόνιων παθήσεων συνιστά μια από τις σημαντικότερες προκλήσεις των σύγχρονων συστημάτων υγείας. Η επιτακτική ανάγκη της συνεχούς διαχείρισης των νοσημάτων αυτών, που συνιστούν αιτία θανάτου για περισσότερο από το 70% του πληθυσμού παγκοσμίως [1], ήταν ένας από τους λόγους που οδήγησαν τον τομέα της ηλεκτρονικής υγείας να γνωρίσει ραγδαία ανάπτυξη. Παράλληλα, η ιδέα της αυτοδιαχείρισης προσωπικών δεδομένων υγείας και τρόπου ζωής, υπό το πρίσμα των νέων τεχνολογιών, κερδίζει έδαφος πολύ γρήγορα. Στις μέρες μας, οι αισθητήρες συνιστούν αναπόσπαστο κομμάτι της καθημερινότητας και συλλέγουν τεράστιες ποσότητες δεδομένων, ελέγχοντας κάθε πτυχή αυτής. Η πρόκληση, λοιπόν, είναι πώς θα καταφέρουμε να διαχειριστούμε όλα αυτά τα δεδομένα που προκύπτουν από το συνδυασμό των υπηρεσιών ηλεκτρονικής υγείας με τις τεχνολογίες φορετών αισθητήρων και κυρίως πώς θα τα ερμηνεύσουμε, ώστε να διευρύνουμε τους ορίζοντες της επιστημονικής έρευνας [2]. Στο σημείο αυτό, ο τομέας της ανάλυσης δεδομένων καλείται να αναλάβει καθοριστικό ρόλο. Οι ασθενείς που χρησιμοποιούν τέτοιες τεχνολογίες, αποκτούν τη δυνατότητα να καταγράψουν και να επεξεργαστούν τα βιοσήματά τους, τις αθλητικές τους δραστηριότητες, τις καθημερινές συνήθειές τους ή ακόμα και τα συναισθήματά τους [3]. Τα δεδομένα που προκύπτουν συνιστούν τον πολύτιμο λίθο της στατιστικής και των τεχνικών μηχανικής μάθησης, η εφαρμογή των οποίων θα οδηγήσει σε εξόρυξη γνώσεων σχετικά με τους παράγοντες αυξημένης επικινδυνότητας για την υγεία ενός ασθενούς και θα παράσχει τη δυνατότητα εξατομικευμένης ιατρικής παρακολούθησης και άμεσης ενημέρωσης για αποφυγή επειγόντων περιστατικών.

Η παρούσα διπλωματική εργασία προτείνει μια μεθοδολογία ανάλυσης δεδομένων που θα εξετάσει τη συνέπεια των ασθενών στο πρόγραμμα λήψης των μετρήσεών τους και θα μελετήσει την αλληλεπίδραση μεταξύ των διαφορετικών ημερήσιων μετρήσεων, με σκοπό τον προσδιορισμό του τρόπου με τον οποίο αυτοί οι παράγοντες μπορούν να επηρεάσουν την παρακολούθηση της υγείας των ασθενών. Παράλληλα, θα πραγματοποιηθούν μελέτες που γενικεύονται σε δημογραφικό επίπεδο, συμπεριλαμβανομένου του φύλου, της ηλικίας και της γεωγραφικής κατανομής, έτσι ώστε να εντοπιστούν οι στατιστικά σημαντικές διαφορές στις ιατρικές τιμές ανα πληθυσμιακή ομάδα και να εξαχθούν πιο στοχευμένα, κατάλληλα συμπεράσματα. Στοχεύοντας στη βελτίωση και εξατομίκευση της ιατρικής παρακολούθησης χρόνιων καταστάσεων υγείας, η προτεινόμενη λύση δύναται να αντιμετωπίσει τις προκλήσεις των ηλεκτρονικών υπηρεσιών υγείας, παρέχοντας στους ασθενείς τη δυνατότητα έγκαιρου εντοπισμού επικίνδυνων καταστάσεων, ενίσχυση της ευημερίας τους, κινητοποίηση για συμμόρφωση στο πρόγραμμα λήψης των μετρήσεών τους αλλά και την εξειδικευμένη θεραπευτική τους αγωγή, δέσμευση για άσκηση και, τέλος, μοντελοποίηση της συμπεριφοράς τους με σκοπό τη βελτίωση της φροντίδας του εαυτού τους και την απόκτηση μιας καλύτερης ποιότητας ζωής.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ανάλυση Δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: διαχείριση χρόνιων νοσημάτων, προσωπικά δεδομένα υγείας και τρόπου ζωής, κινητή υγεία, στατιστική ανάλυση δεδομένων, μηχανική μάθηση

ACKNOWLEDGMENTS

The present thesis has been developed in collaboration with the Software and Knowledge Engineering Laboratory of the Institute of Informatics and Telecommunications of NCSR “Demokritos” and the company BioAssist S.A. and I would like to express my whole – hearted gratitude to those who have helped me to complete it.

First and foremost, I would like to thank my supervisor, Post-Doctoral Research Associate at the SKEL – IIT, NCSR “Democritos”, Dr. Anastasia Krithara, whose constant encouragement and support followed me in every step of this study. Her positive attitude and her prompt and firm response to all of my problems throughout the research have been invaluable, while her comments and scholarly advice have been a true source of inspiration for me. Without her continuous support and guidance this thesis would have never been completed.

Of course, the present thesis could not become reality if it weren’t Dr. Ilias Maglogiannis, Associate Professor at the Department of Digital Systems, University of Piraeus, a person I have the honor to consider as my mentor. I sincerely thank him for believing in me from the beginning and, through BioAssist S.A., giving me the unique opportunity to work in an area that corresponds perfectly to my scientific interests. I also thank him for his confidence in me by entrusting me to deal with such an interesting topic as my thesis project, whose completion would not be possible without his valuable support.

I would also like to thank Dr. Vangelis Karkaletsis, Research Director of the Institute of Informatics and Telecommunications, Head of the Software and Knowledge Engineering Laboratory, NCSR "Demokritos", for the valuable time he devoted to my thesis and for honoring me as a member of the three-member examination committee.

I owe special thanks to Dr. Andreas Menychtas, Post-Doctoral Research Associate at the Department of Digital Systems, University of Piraeus, and CTO of BioAssist S.A. for his generous help and the guidance he provided me about the data and the most technical aspects of the thesis.

I also take the opportunity to thank my colleagues and friends for their constant support and encouragement throughout these years.

Last but not least, I owe a special dept to my family for their continuous mental and moral support and encouragement. My mother, my father, my brother, my grandmother and my grandfather, never stopped believing in me and standing by me, assistants in all my goals and dreams, invaluable support and a unique part of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INSTEAD OF PREFACE

Sapias, vina liques, et spatio brevi
spem longam reseces. Dum loquimur, fugerit invida
aetas: carpe diem quam minimum credula postero.

~ Horatius,
Carmina 1.11, 6-8

1. INTRODUCTION

1.1 Problem Definition

The public health issue of chronic diseases constitutes the global social scourge of our times, presenting the highest rates of morbidity and mortality [1]. The only way to handle this crucial and urgent situation is through early detection and management of potentially dangerous incidents which can be achieved effectively if the patients are capable of self-monitoring the condition of their own health.

The need for constant monitoring of chronic health situations has led scientists to explore three main fields of technology: “Internet of Things”, “Mobile Health” and the “Quantified Self Movement”. The combination of these three principal axes results in “Big Data” regarding a patient’s condition in relation to aspects of their daily life. In order to gain knowledge from these data and be able to personalize the medical follow-up and immediate feedback of patients, statistical and machine learning methodologies is imperative to be applied. Data analysis will extract the maximum possible information in order to answer questions such as “why” a patient is in such a medical situation or “how” they can manage their condition on their own in the future [4].

The incorporation of these techniques into modern electronic health services, can improve the welfare and quality of life of chronic patients. Providing them with the appropriate information in the most comprehensible way, data analytics play a decisive role in the modeling of the patients’ behavior, motivating them to abide by the medical instructions and to engage in physical activities, thus, enhancing self-care.

1.2 Thesis Contribution

The present thesis project is engaged in an attempt to present the trends and needs of our times, both scientifically and technically and – based on a specific case study – propose a data analytics solution that could control the data, understand them and drive to conclusions that could be used as valuable feedback both to the stakeholders and the chronic patients, in order to enhance the welfare and longevity of the latter.

For this objective, statistical and machine learning methodologies are being applied to data concerning the biomedical and activity information, along with the demographic characteristics of patients, as gathered from various IoT devices. The goal is to examine patients’ consistency in their measurement schedule and to study the quality of the measurements in an everyday basis along with the interaction among them, in order to determine if there is an impact on the monitoring of patients’ health condition, to early detect possible dangerous situations and, thus, to enable self-monitoring. Last but not least, the results are generalized on a sex, age and geolocation level so that appropriate recommendations can be derived per population.

For the needs of the analysis, data concerning patients’ demographic, biomedical and activity information are provided by the company “BioAssist SA”, as gathered from various IoT devices in the context of the “HeartAround” mobile service.

1.3 Thesis Outline

The current thesis project is organized as follows: Chapter 2 constitutes a research review and provides the necessary scientific background on chronic diseases, the Quantified Self Concept and mobile Health. Chapter 3 focuses on the technical background of the Big Data analytics process. Chapter 4 presents the case study whom data are used for the current implementation. Chapter 5 describes the proposed analytics solution, along with the experiments performed and their results. Chapter 6 concentrates on the languages, the libraries and the tools that were used for the proposed implementation and finally, yet importantly, Chapter 7 provides the conclusions of this work and suggests the potential future scientific extensions that can follow.

2. SCIENTIFIC BACKGROUND

2.1 Chronic Diseases

Chronic diseases, also referred to as non-communicable diseases (hereafter, NCDs) constitute a decisive health factor, reaching the highest morbidity and mortality rates globally. Due to their increasingly rapid prevalence in human health over the years, they are one of the greatest challenges in the field of healthcare.

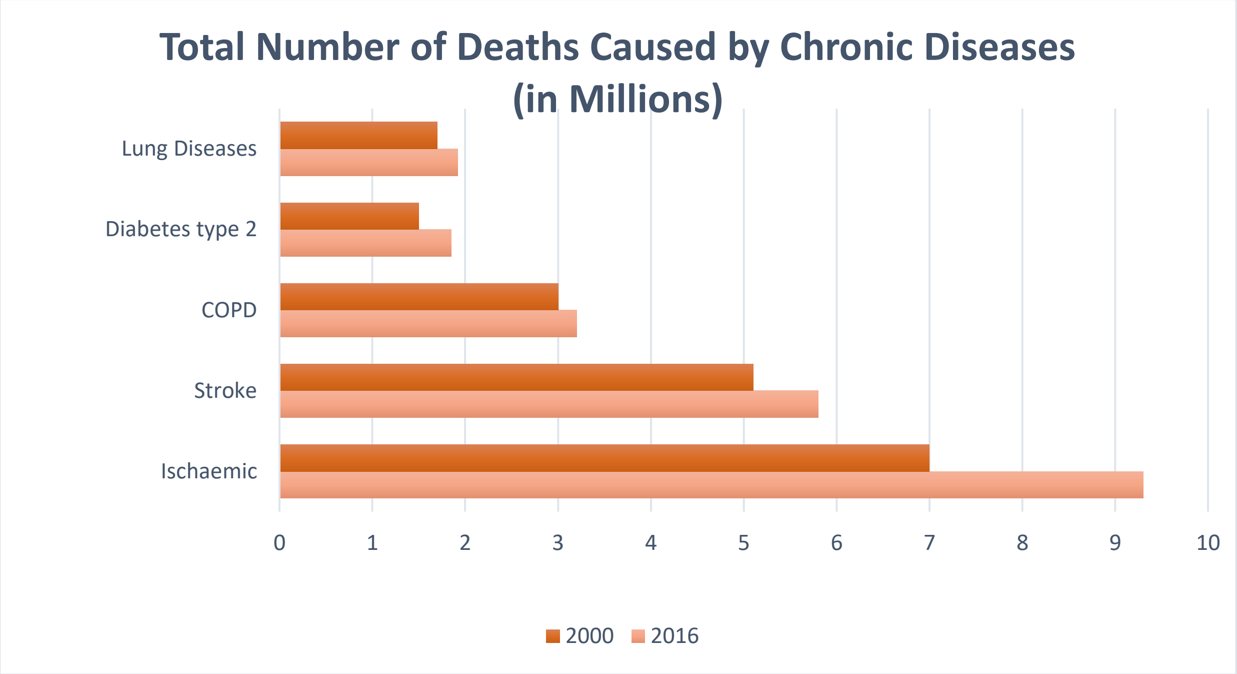


Figure 1: Increase in the No. of Deaths Caused by Chronic Diseases 2000 – 2016 [5]

2.1.1 Risk Factors

According to the World Health Organization (WHO) [6], there are four main causes that drive the phenomenon of the dominance of chronic diseases, allowing them to reach the levels of epidemy, both in the developed and the developing world: unhealthy diet, uncontrollable consumption of alcohol, physical inactivity and tobacco use. The repercussions of these behavioral risk factors are the gradual development of cardiovascular diseases, chronic respiratory diseases, diabetes mellitus and cancer, among others.

Table 1: Risk Factors Causing Chronic Diseases [6]

	Unhealthy Diet	Uncontrollable Alcohol Consumption	Physical Inactivity	Tobacco
Cardiovascular Diseases	X	X	X	X
Respiratory Diseases	X	-	-	-
Diabetes	X	X	X	X
Cancer	X	X	X	X

Table 1, above, and Figure 2, below, represent the correlation of the aforementioned factors to these severe diseases. Only if these risk factors were somehow reduced, which would mean early detection and timely treatment, a significant diminution of the proportions of this phenomenon could take place [7]. In this context, it goes without saying that it is essential to confront this challenge.

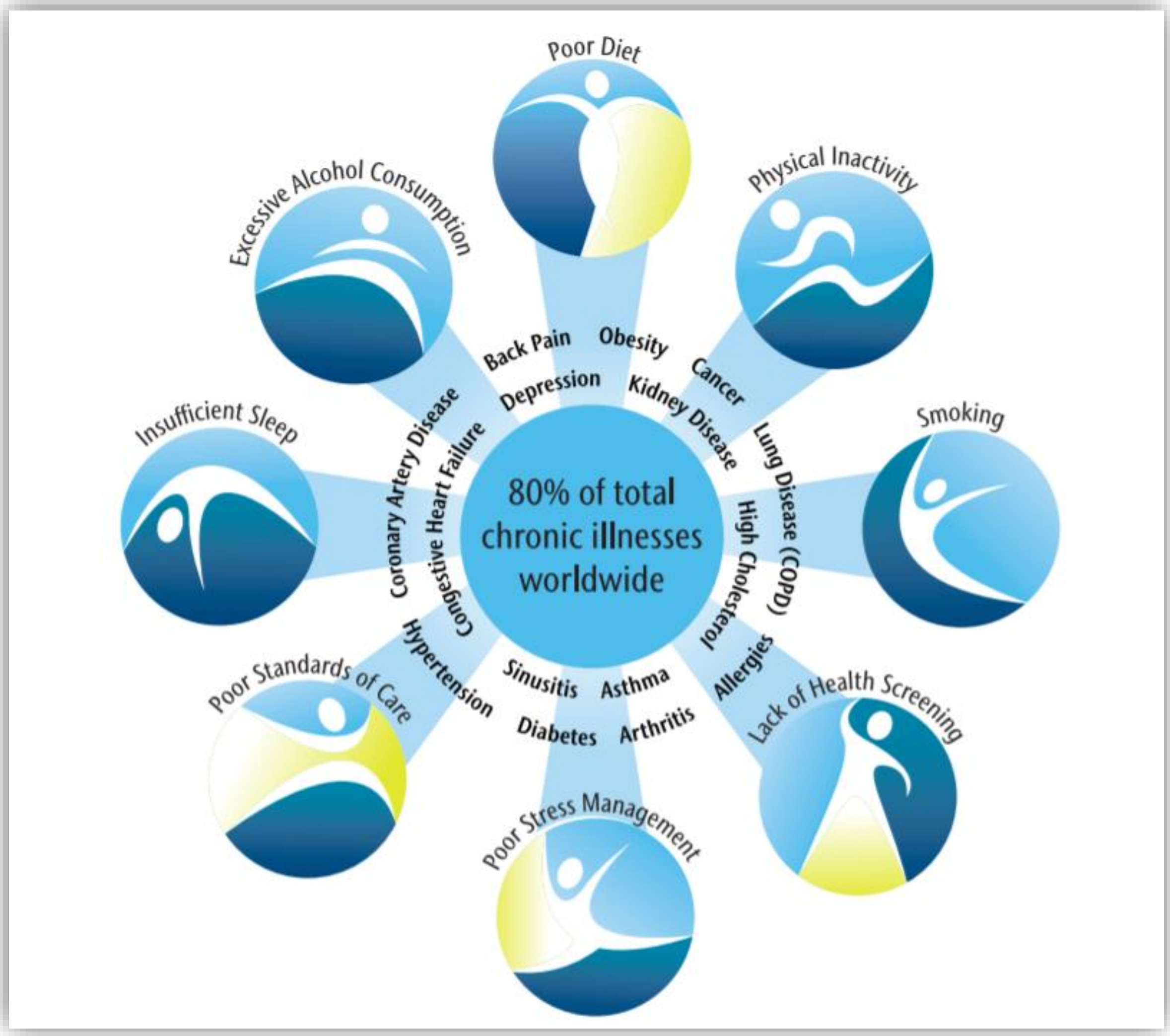


Figure 2: Causes of Chronic Illnesses Worldwide [8]

2.1.2 Chronic Diseases Management

Nowadays, not only have chronic diseases increased in variety, but also there has been a noticeable augmentation in the number of people suffering from them during the last decades, as shown in the histogram of Figure 3.

It is commonly accepted that in order to manage the prevalence of chronic diseases, promoting health is not enough. Risk prevention as part of healthcare is what should be of the highest priority. More specifically, disease prevention is discretized in primary, which is linked to disease prolepsis (i.e. before it occurs), secondary, which involves screening and attempts to understand the real factors that are responsible for the appearance of the disease and to identify it in the earliest possible stage so that its progress can be restrained, and last but not least, tertiary, which focuses on the patient’s treatment and finding the quickest and less painful recovery.

Self-management of a chronic disease can cover all the tree types of prevention, since the patient is able to understand the disease and live with it, while controlling their condition and maintaining a better quality of life.

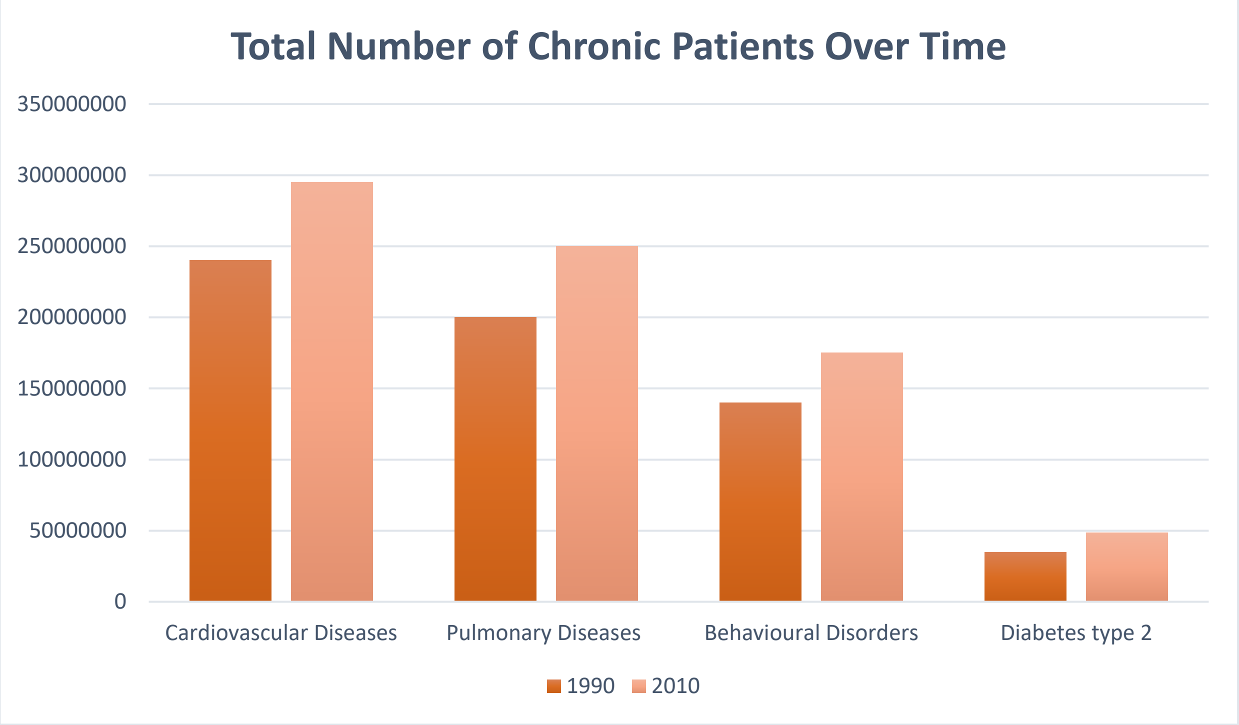


Figure 3: Increase in the No. of Chronic Patients 1990 – 2010 [9]

Studies of the chronic care model [10] prove that the effectiveness of healthcare improvement is a multi-factor function. They underline that our focus should target:

- The changes that need to be performed in healthcare systems, so that their processes can be sufficiently developed,
- The support for self-management that should be provided to the patients, so that they can utilize the maximum benefits,
- The adequate guidance that should be provided to the healthcare practitioners in order to improve their collaboration with their patients, and, finally and most importantly,
- The development and improvement of modern Information Systems so that they can be incorporated and play a decisive role in the care model.

The same studies point out that for a healthcare system to be improved and effective, the self-management component should be standalone and sufficient. The importance of this idea makes it obvious that self-management is not easily applicable, yet it is of imperative need to be performed, as it can define the progress of a chronic patient’s health situation.

2.1.3 Vital Sign Measures

The most common types of data that a chronic patient records can be divided into two basic categories. The first one is related to the biosignals of the patient, the continually measured and monitored signals that are produced by each electrical, chemical and mechanical activities that appear during a biological event [11], and the second one to the patient’s physical activity data.

Depending on the chronic disease that a patient suffers from, the measures they need to monitor are usually the following:

- **Oxygen Saturation:** An indicator that reflects the percentage of hemoglobin that is saturated with oxygen, in relation to the total amount of hemoglobin present in the blood.
- **Blood Pressure:** The force that promotes the blood through the arteries to all the tissues of the body, ensuring its continuous circulation. It is expressed by two partial values: systolic and diastolic blood pressure. The systolic one corresponds to the pressure that is exerted on the arteries when the heart contracts and feeds them with blood. The diastolic pressure refers to the one exerted on the arteries when the heart calms after contraction. The unit of measurement is mmHg (i.e. millimeters of mercury).
- **Heart Rate:** Heart rate, or pulse, is the transmission of the blood wave, that is caused by the heart's contraction to the vessels. The pulse is the result of the change in blood pressure caused by abdominal contractions of the heart and elasticity of the arteries. The unit of measurement is the bpm (i.e. beats per minute).
- **Glucose Levels:** The blood glucose concentration, or blood sugar level, is the amount of glucose that is present in the blood. Glucose is a simple monosaccharide found in plants and constitutes one of the three dietary monosaccharides, along with fructose and galactose, which are absorbed directly into the bloodstream during digestion. It is the most important carbohydrate, since cells use it as their primary source of energy and a means of metabolism. The unit of measurement is mg/dL (i.e. milligrams per deciliter).
- **Spirometry Values:** The measuring of breath, or the pulmonary function test values, measures the function of the lungs, and more precisely the volume and flow of the air that a person inhales or exhales. The unit of measurement depends on the parameters measured. Some of the most common are: Forced Expiratory Volume (FEV) measured in L, Forced Vital Capacity (FVC) measured in L, Peak Expiratory Flow (PEF) measured in L/m, FEV1_FVC ratio measured in % and the Forced Expiratory Flow (FEF2575) measured in L/s.
- **Weight:** Refers to the human body weight and constituted the measure of mass. The most common unit of measurement is kg (i.e. kilogram)
- **Activity:** A patient's physical activity, such as the type of activity (walking, running, sleeping etc.), the duration (minutes per day) and parameters for assessing the physical activity such as the burning of calories (kcal/day).

2.2 Quantified Self

Quantified Self (hereafter, QS) is the concept of individuals who exploit the new means of technology to self-monitor their life by collecting and storing their own medical or life data, applying analytics methodologies in order to achieve the most optimum result, which is to understand the reasons that led them in this particular health condition [4]. Pressingly, technologies like smartphones, mobile applications and wearable sensor devices give the potential to individuals to learn and control many different aspects of their daily life, varying between vital signs such as heart-rate, blood pressure, or blood glucose levels, and lifestyle quality such as mood, sleep and activity [12].

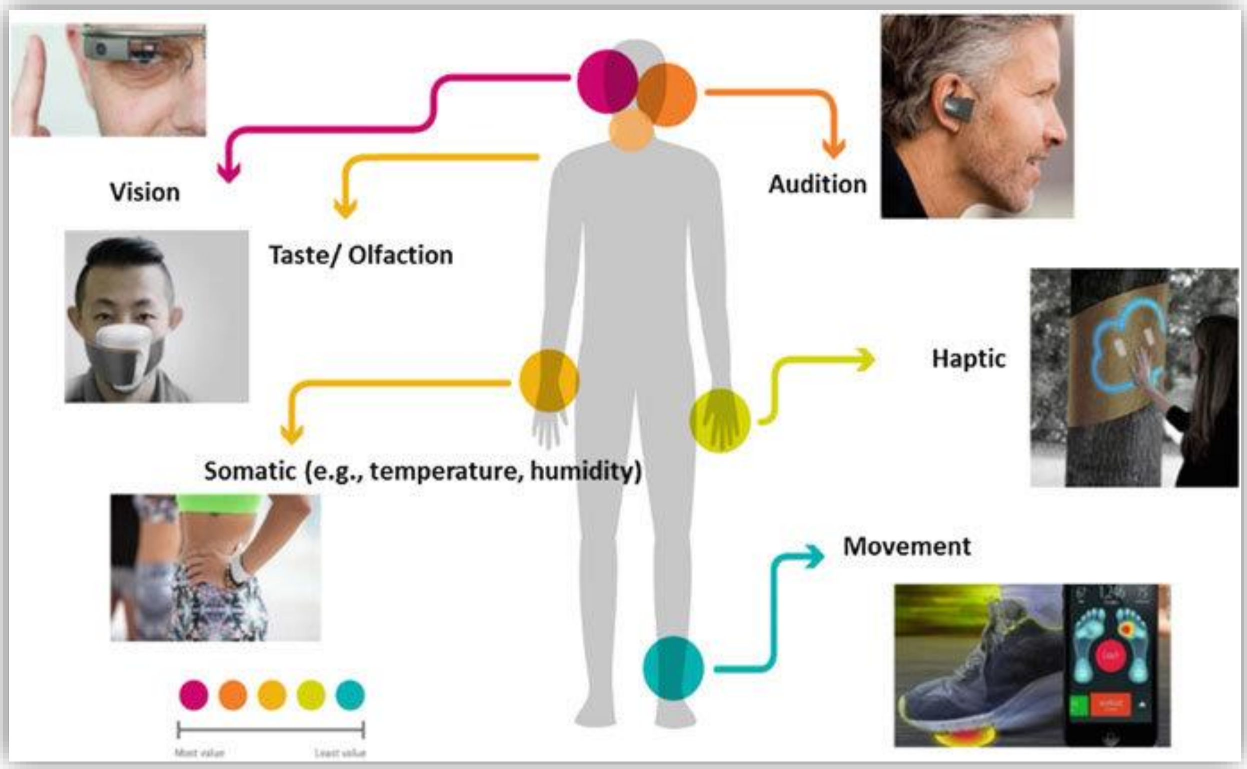


Figure 4: Quantified Self [75]

Within this framework, the deployed QS sensors and monitoring devices gather the aforementioned data and allow the individuals to explore themselves and improve their quality of life.

2.2.1 The Movement

The term “Quantified Self” was first conceived in 2007 by Gary Wolf and Kevin Kelly [13] and, as established later by the homonymous community, reflects the idea of ”self knowledge through numbers” [14]. These two editors of America’s Wired magazine had the vision of helping people to understand the meaning of their own personal data. In this context, individuals all over the world started to self-capture data about their life, over a period of time, using mobile devices and wearable technologies, with the aim of analyzing them and expanding their knowledge regarding the “self”. The movement of “lifelogging”, as it is alternatively called, focuses on the incorporation of new technologies into the different aspects of an individual’s life. Data acquisition of the hemoglobin saturation levels in the blood, systolic and diastolic blood pressure, heartbeat, glucose levels and many other biological signals measured by self-sensing devices, along with performance data, such as activity, sleep duration and quality or mood, are some of the interests of the QS community users and the QS tracking tool makers [15]. As a next step, analyzation techniques are being performed, in order to make sense about the meaning of humanity.

Based on the above, it can thus be said that the Quantified Self Movement constitutes the “leading edge” of the world of Internet of Things (hereafter, IoT) [16].

2.2.2 Evolution

The past decade has been characterized by an explosion of interest in self-quantifying. However, this phenomenon is not a novelty of the 21st century. Athletes have been monitoring their health and activity, such as calorie intake, weight or performance, using scales or time meters, for many decades already. For years, people have been using diaries to record their feelings, mood or experiences. What is currently considered to be an innovation is the promptitude with which the progressive transition from the old habits to the QS movement became a trend and spread widely, as well as the extent to which new technologies are supporting it.

This ever-increasing tendency is the result of the following components:

- **Smart Devices:** Smartphones and tablets constitute an integral part of an individual’s everyday life. Equipped with built-in sensors, such as GPS or three-axis accelerometers, they are able to capture data about the location or the activity and motion of the user (respectively) and, thanks to cloud computing, they are also able to transfer, store, manage and process these data very easily.
- **Wearable/ Sensor Technologies**: Becoming increasingly popular as a considerable lifestyle trend [16], fitness oriented wearable devices and body sensors, along with medical devices, that constantly monitor metrics such as an individual’s steps, weight and sleep patterns, or blood glucose levels, oxygen saturation, blood pressure, etc., are gradually becoming increasingly powerful, as well as easier to transport and to use.
- **Social Media:** More and more people share their personal data online, providing information to their friends, or even to everyone publicly, about their life habits, new hobbies, performances, or feelings [17].

The evolution of QS is attributable to the above factors, the development of which has attracted interest in the creation of future applications for mobile and wearable environments.

2.2.3 Indications of Growth

Taking into consideration the above three factors that assisted in the rise of the Quantified Self field, one can safely expect that the data that are being collected are not only innumerable, but also constantly increasing. This phenomenon results in making users increasingly demanding. Since the collected data grow exponentially, users can become unable to understand their meaning by themselves. The more data are being tracked, the more requirements and needs derive. Advances in the field of Information Technology (hereafter, IT) lead to technological advances that bring these, considerably unapproachable, data of health and fitness within the users’ reach. At the same time, the tendency of aggregating the data is rapidly increasing as well. By merging the different types of data of an individual, or of many individuals, a more complete image is being produced. In this context, it is expected that data from various sensor devices could potentially be shared and combined, so that their analysis can achieve additional dimensions in the framework of personal knowledge discovery, or regarding the in-depth knowledge of a population [16]. As Honan noted in the Wired Magazine [18], ‘‘My big hope for the next generation of health-tracking apps is that they will help us understand what things mean and give us the tools to act-instead of just numbers’’.

2.2.4 Fields of Interest

In this day and age, more and more projects are developed in order to track and acquire knowledge from data, related to the different aspects of an individual's life. The common denominator of the different applications, services, devices and tools used for the aforementioned objective, is summarized in Table 2, which reflects the principal fields that attract the interest of QS [19].

Table 2: Quantified Self Fields of Interest

Fields of Interest
Healthcare
Fitness
Mood

- **Healthcare:** One of the most essential domains related to a human's life is healthcare. The need for developing smarter and more accurate and efficient sensors, as well as services to satisfy the ever-changing requirements of this field is ever-growing. The self-tracked data from each device provide a varied view of an individual's different vital signs. The involved users are able to statistically analyze the various factors that influence their biosignals, and potentially find a way to reclaim and reinforce their health status.
- **Fitness:** Tracking data regarding the everyday physical activity and performance, the weight equibalance, or the consumption of calories, in order to enhance self-awareness and a healthy lifestyle in the area of athletics and fitness, is an increasingly common task. Nowadays, the development of wearable devices, such as sensor activity trackers, is a rapidly expanding trend. Some of these devices count the steps or measure the quality of sleep, using three-axis accelerometers, while others exploit also GPS technologies to give the user the possibility to track the duration of the activity, the speed, the distance travelled, or the calories burned, with the aim to provide them with an overview of their progress in real-time. Simultaneously, such data related to smartphone applications, can be integrated with other similar services, shared and then correlated and compared to those of other users.
- **Mood:** What is considered to be of particular interest is the area of mood tracking. Mobile applications try to identify changes in the mood and behavior of the user over a period of time, in relation to environmental or psychosocial factors. Devices and services intend to make users understand the exogenous parameters that have an impact in their mood and mental condition. This can be achieved using many different ways, such as gamification, camera emotion captures, or even manual input by the user [20].

2.2.5 Security Concerns

This burst of information, has undeniably given birth to concerns [21] that QS data could be intercepted and used for negative purposes. For this reason, companies give every day more and more focus on enhancing the security issues of their devices.

2.3 Mobile Health

The newest trend of our times, that more and more individuals use, are the increasingly popular smart devices and wearable technologies that monitor and manage their health conditions. This has undoubtedly given rise to a new model of practicing medicine and public health: the mobile health (hereafter, mHealth) [22]. According to the Global Observatory for eHealth (GOe), mHealth, component of eHealth, is defined as the “medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants and other wireless devices” [23].

The utilization growth of this emerging field of mobile technologies in healthcare, has particular significance in improving the lifestyle behaviors of patients, ultimately reducing the risk for many chronic diseases, and can also contribute to either facilitating the communication between patients and medical practitioners, or provide patients with their medical information without the involvement of their doctors.

Figure 5: Mobile Health [24]

2.3.1 Big Data

In the last few years, there has been a noticeable, exponential augmentation in the quantities of the produced digital information. The term “Big Data” has been used to describe the – structured or unstructured – data, the volume of which exceeds the capabilities of the commonly used software to record, manage and process them in a 'sustainable' time period. Data collection has been an extremely rapid procedure, executed from relatively inexpensive devices/ things that have access to the internet (Internet of Things), such as smartphones, sensors, software logs and cameras, making the conventional relational databases unable to manage them.

The widespread usage of Big Data could constitute the cornerstone of a healthier lifestyle, providing effective tools for changing people’s mentalities and attitudes. In particular, the use of mobile phones in health (mHealth) has the ability to tailor these tools and personalize them for each individual, by using lifestyle data, such as nutrition, physical activity, sleep, medical condition, etc. and assessing the impact data in individual population groups. In addition to providing information about people, mHealth technologies exploit this relevant information, which is the key to innovation and, as a result, they provide a fully comprehensive picture of what influences a patient's progress and what causes the setbacks in the provided treatment. Thus, Big Data, through mHealth and its technologies, can help improve the early detection of diseases, and contribute to the adequate management of negative health factors.

2.3.2 Internet of Things

In 2008, the number of Internet-enabled devices exceeded the number of people on the Internet and is estimated to reach the number of 50 billion [25] or more by 2020. And the term devices does not only refer to the traditional computers or laptops, smartphones and tablets, but it also includes all the different kinds of the so-called "smart" objects in the natural world, from sensors and home appliances to cars and entire buildings.

This enormous explosion of the variety of devices, in number and specification, that, in this day and age, have the ability to be connected to each other and to the Internet, has led to the development of the "Internet of Things" ecosystem. Hence, the Internet of Things (IoT) can be defined as the interconnection of physical objects, vehicles, buildings and, in general, systems that incorporate a computing system, so that these objects can communicate with each other, exchange and process data [2].

Figure 6: The IoT Ecosystem [20]

Among these various devices, those that are of significant importance, and are studied in the context of the current thesis project, are the "smart" medical devices, such as oximeters, blood pressure monitors, scales, etc. These devices constitute the newest variants of the conventional medical instruments and while they receive biosignals measurements, they have the simultaneous ability to connect to computers or smart phones and tablets and transmit the metrics in real time. This is usually achieved via Bluetooth Low Energy (BLE) wireless technology and among devices that support Bluetooth connections [26]. In the same context, activity tracking systems shall be included. Smart mobile devices' applications can monitor and record the user's/ patient's daily routine and physical activity, such as sports, other activities, eating habits, calorie consumption and many other types of data that can be quantified, thanks to the special bracelet sensors (linked via BLE), as the next sections will present.

2.3.3 mHealth Services

The technologies of recent years have helped, not only to obtain data from the healthcare environment, like hospitals, health centers or laboratories, but also directly from the patients through the sensors, their monitoring, the Internet of Things, etc. Health services will benefit directly through the acquisition and analysis of information from all these different sources.

Ideally, health services should not be restricted to hospitals and clinics but should and must be incorporated to home care services, especially in cases of elderly chronic patients. Mobile health devices support the real-time monitoring of the patient through smart mobile phones ore tablets and wearable, wireless medical instruments and, their medical Big Data such as blood pressure, blood glucose levels, oxygen saturation and so on, can be collected simultaneously, stored to cloud services and become available to their attending physicians 24/7. In this way, the patient stays in supervision all the time as their monitoring successfully becomes remote. This potential, except for its obvious economic impact, can affect the psychology of the patients, who most of the times prefer to be in their own, familiar environment, instead of unfriendly hospitals, because it helps them feel safer, knowing that there is a vigilant eye “watching”, and lets them have self-control of their health, which is proven to be related to the improvement of their psychological situation.

Figure 7: IoT Adoption in Healthcare – “Internet of Medicine” [21], [23]

2.3.4 Technologies and Sensor Devices

Among the various technologies and devices that can be incorporated to a mobile health service, the most common ones are the following:

Pulse Oximeter: Pulse oximetry calculates the saturation of oxygen with the use of light. The emitted light (from light sources), crosses the pulse oximeter’s catheter and reaches the light detector. If a finger is placed in-between the light source and the light detector, the light will have to pass through the finger to reach the detector. During this procedure, part of the light will be absorbed by the finger and, therefore, only the portion of light that is not absorbed will reach the detector. The amount of light absorbed by the finger depends on many physical properties and these properties are used by the pulse oximeter to calculate the oxygen's saturation. During the analysis of the pulsed arterial blood, the pulse oximeter has the ability to present the changing portion of the light's absorbance in a graphical form. This is called Photoplethysmogram (PPG) and constitutes an extremely important graph to be studied, as it can assess the quality of the pulse signal.

Blood Pressure Monitor: The sphygmomanometer device that is used to measure the blood pressure. It consists of two parts, an inflatable cuff and an electrical mechanical manometer that measures the pressure [27]. More specifically, the cuff is placed over the upper arm or wrist of the patient and inflates until it reaches a pressure of around 20 mm Hg above the patient's systolic one, when, there is no blood flow through the patient's artery. Then the cuff starts to deflate slowly and the reducing pressure that is exerted on the artery allows the blood flow again and, as a result, vibrations start to be detected on the arterial wall, since the blood has to push the wall to open in order to flow through the artery. When the pressure that the cuff exerts falls below the patient's diastolic one, the blood is able to flow smoothly through the artery, in its usual pulses, causing no more vibrations to the arterial wall. These vibrations are the measurements that are transferred to the monitor through the air in the cuff and get converted to electrical signals [28].

Glucometer: Glucose meters provide readings by detecting the approximate concentration of glucose in a person's blood. To get a reading, the skin needs to be pricked – most commonly, a finger – and the drop of blood that is acquired needs to be applied, as a blood sample, to a disposable test strip that is inserted in the meter. The glucose in the blood reacts with the chemicals in the strip and then, electrical currents pass through and determine the levels of glucose in the blood sample, providing numerical results only within a few seconds [29].

Spirometer: The apparatus that takes measurements of the lung's inhaled and exhaled quantities of air during a specific time period, in order to determine a person's respiratory capacities. Practically, when the patient breaths through the spirometer's hose, the pressure sensor that is located in its interior converts the flow of air in an electrical signal which is later processed by electronic means and, as a result, gives measures of the flow and calculations regarding the spirometry parameters [30].

Scale: The dynamometer device that measures the vertical force exerted by the body that is located on them (i.e. the weight). The result of the measurement, however, is not in units of force (Newton), but in units of mass measurement (kg). Essentially, they show the mass which, in the gravitational field of the earth, weighs the force exerted on them [31].

Activity Tracker: A wearable fitness tracking device that monitors and tracks metrics related to the physical condition of the patient, thanks to the different types of sensors that can be incorporated into them, including 3 axis accelerometers that tracks the movement in any direction, optical sensors that use the light on the skin to measure the pulse, temperature sensors to keep track of the temperature changes etc. [32].

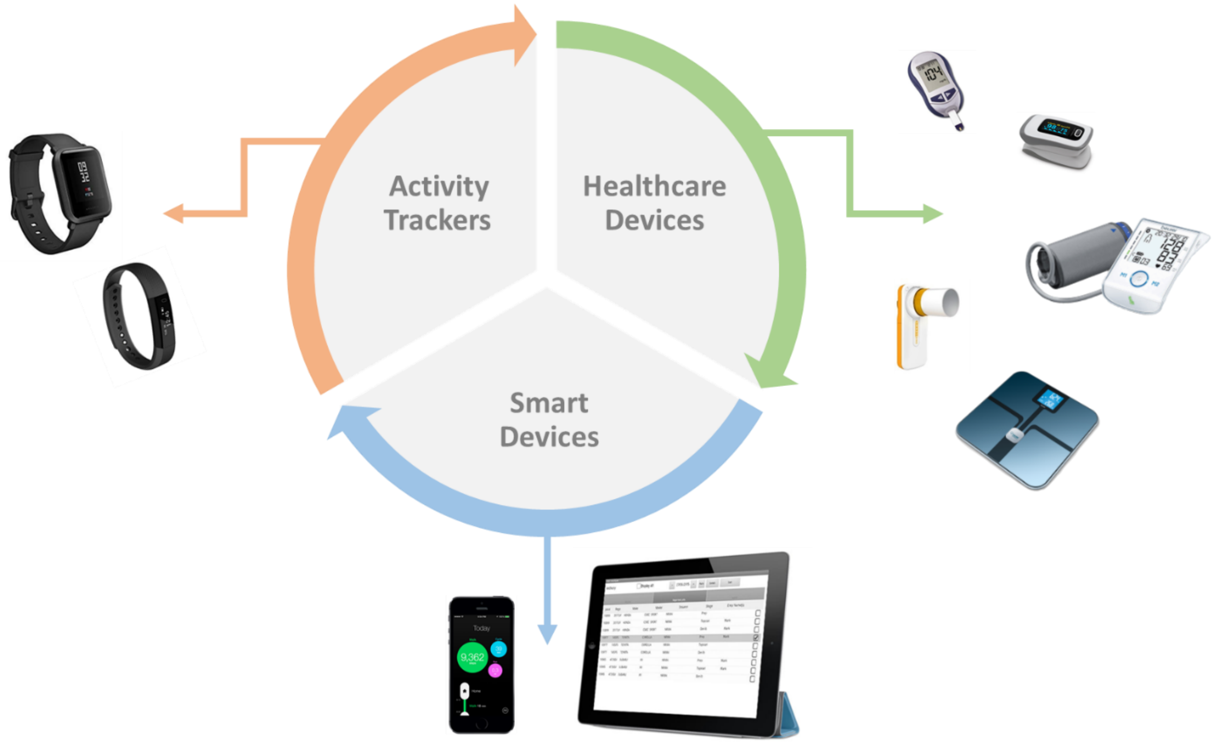


Figure 8: IoT and mHealth [33]

In light of the literature review, it is obvious that chronic diseases, the Quantified Self movement and the revolution of the new, smart technologies are interrelated. What is worthwhile at this point is to see how all this information can be exploited, by analyzing it, and to draw safe conclusions about the patients' future.

3. TECHNICAL BACKGROUND

People have the natural tendency to leave traces behind; sometimes on purpose, and others unconsciously. Smart devices and new technologies have undoubtedly contributed, to the generation of enormous amounts of incomprehensible “Big Data” out of people’s everyday life, as described in the previous chapter. Being able to understand the complexity of these data and make sense out of them, requires familiarization with the technical aspects and methods that can lead to answering the “how”, “who”, “when” and “what” questions that derive from their analysis. This chapter focuses in giving an insight of the main steps that an analysis procedure follows, along with the most common and effective methods that contribute to its implementation.

3.1 Data Analytics Methodology

Data analysis covers a wide area that is constantly changing and is extremely complex. The different methods, architectures and tools that are utilized for the analysis purposes, produce different information. Traditional data analysis requires appropriate statistical methods, in order to maximize the value of the information. Data analysis plays an important role in decision-making, understanding them, discovering knowledge and forecasting potentially dangerous situations and these are some of the reasons why it is vital to be applied.

As Big Data are not a simple structure, it is essential that several steps be performed to analyze them and make them useful [34]. A general overview can be presented in the following figure:

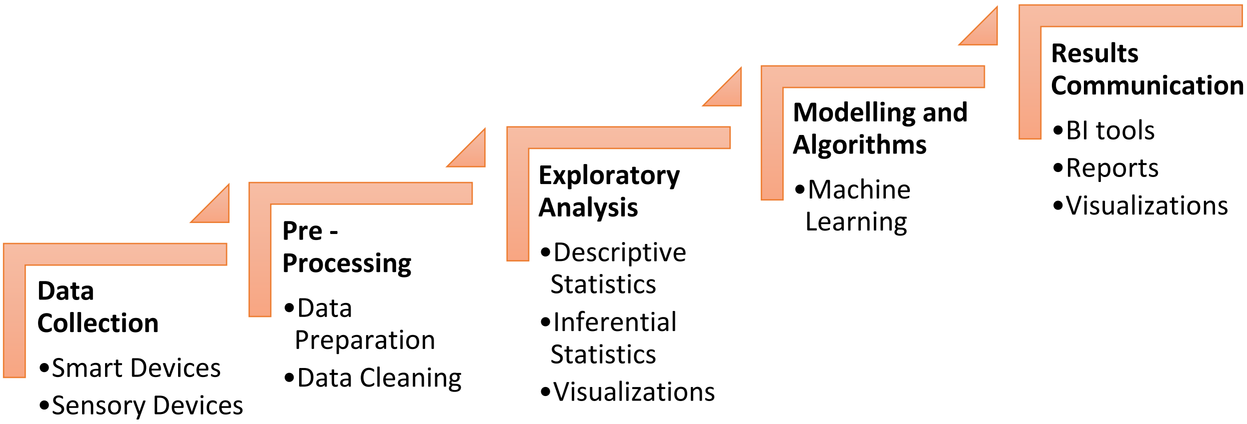


Figure 9: Steps of Big Data Analysis

Starting from the recording and the collection of data, the process moves into their preparation and cleaning so that they get transformed in an appropriate format, in order to be explored with statistical approaches, followed by their visualizations. The completion of the “data understanding”, leads the analyst to select and extract the most suitable features to be analyzed. This is where statistical modelling and machine learning algorithms take place and perform the analysis of the former-complex Big Data. Last but not least, the final level, yet the most important, is the interpretation of the results and their communication.

3.2 Data Collection

The data can be generated by various means and environments. Indispensable prerequisite of their analysis is to be in the analyst’s disposal. So, the “data collection” process includes the information acquisition from all the relevant sources, i.e. the sensory technologies and the mobile equipment, as it has already been explained thoroughly in the previous chapter.

3.3 Pre-Processing

The wide diversity in sources of data engages the analyst to tackle problems such as noise, redundancy or deficiency of data and inconsistency, among others. There is no denying the fact that the quality of the data certifies the accuracy of the analysis, so, the more prepared the data are, the better and more precise results will be reached, otherwise, the research will be misled and result in false conclusions. Therefore, in order to achieve effective quantitative data analysis, several actions need to be performed primarily.

Some of the most common approaches that it is necessary to be performed are summarized as follows:

3.3.1 Data Preparation

Data cleansing is the most critical part of the analysis. Preparing, or formatting, the data means defining and correcting mistakes such as error types, transforming their format and modifying or deleting the information that is inaccurate, nonsense, incomplete or needless. Through this step, the vital need of preserving the data's consistency can be achieved [35].

3.3.2 Detecting Outliers

When working with sensory data, the control of extreme values, which are most probably likely to happen, is inevitable. These values are the ones called outliers and constitute observations that are located in abnormal distance from the others in the examined sample. What is tricky about outliers is that, although in their vast majority constitute aberrations as a result of measurement errors, they could also contain valuable information in case they have been caused by the alterability of the occurrence that is being observed [2]. Consequently, it goes without saying that it is of great importance to detect them and understand their physical meaning so that the research results maintain their validity.

A popular method to achieve the outliers' detection is the Chauvenet's criterion, which creates a band of data around the mean and identifies the values that are located outside of it [36]. Among other techniques that fulfil this purpose, are the mixture model-based or simple distance-based calculations [2], the local outlier factor, which identifies the local density of the attributes and compares it to the others [2] and z-score. The latter is the one that is mainly implemented in the context of this project.

Z-Score, follows the same logic with the Chauvenet's criterion and attempts to describe a data point based on the relationship it has with the mean and the standard deviation of the group of the examined data points. In other words, z-score signifies the mapping of the data to a normal distribution and, once they are centered and rescaled, those that are far from z-score thresholds, $-3 \leq z \leq +3$, are considered to be outliers [37]. Further information on this technique is presented in the Data Standardization section.

3.3.3 Handling Missing Values

A common problem in data preparation is when some data are missing and there is no other way to acquire them. When coping with internet of things, this phenomenon is inevitable because of the intrinsic characteristics of the data and its occurrence, which leads to data incompleteness, can be provoked by multiple reasons, such as network issues that could result in synchronization failures, asynchronous data transmission from the devices, deceitful devices or device malfunctions, etc. [38]. Handling of these missing data is essential, as most of the algorithms cannot ignore them and decide to replace them automatically, misleading the final results, or they are not able to function at all. At this point the analyst should take place and perform the most appropriate missing value handling technique, in order to have the analysis under control.

One of the most useful methods is the imputation of the mean that could be used to replace the missing values. A very useful technique to be performed for this purpose is the k - Nearest Neighbor algorithm (hereafter, k-NN) [39] that attempts to match a point with a missing value, with its closest k - neighbors in a multi-dimensional space. The k-NN algorithm will be discussed in the Supervised Learning section.

Among the various other methods that are used for missing values handling, some very usual are the imputation of the median or the mode value, the replacement of missing values with the sample's average value, the extrapolation of the previous and the next measurement's value, the model-based imputation that predicts the missing value, the conservation of the standard deviation of the data or the Kalman filter, that takes into consideration historical observations [2].

3.3.4 Handling Noisy Data

Several times, during the preparation of the data, it is essential that they be smoothened to eliminate the noise and the extreme values. What is considered as noise every time, it is defined by the analyst, along with the filtering that the data will undergo. Data transformation, in terms of noise handling and filtering, can be effectively performed by reducing the dimensionality of the data. Reducing the number of random variables, the dimensions of the features' space decreases and, the fewer relationships among the features the better data exploration and visualization can be performed.

Two of the most useful techniques used for dimensionality reduction and implemented in the context of the current thesis project are the Principal Component Analysis (hereafter, PCA) and the t-Distributed Stochastic Neighbor Embedding (hereafter, t-SNE). More specifically, PCA refers to the whole dataset and transforms the correlated variables into new variables that are uncorrelated, called "principal components", or main axes [2]. This method allows the analyst to reduce the dimension of the variables and, hopefully, remove the noisy ones [40]. Mathematically, it is about a linear feature extraction method that maps the data to a lower-dimensional space where the variance gets maximized. This is the result of the calculation of the largest eigenvectors from the covariance matrix, the principal components, which are used to retain the valuable part of the features of the initial dataset, as they can reconstruct the valuable fractions of their variance [41].

On the contrary, t-SNE, is a non-linear method and it is mainly used for high-dimensional data exploration and visualizations. Its goal is to minimize the Kullback-Leibler divergence between the joint probabilities of the high-dimensional input data and the data points in the low-dimensional space [41]. As a result, the high-dimensional data are mapped to this lower-dimensional space and, based on the similarity of the data points, the algorithm identifies clusters in order to find patterns in the represented data.

The idea of dimensionality reduction [42] can also be performed by many different filtering approaches, some of which are the Missing Values Ratio, the Low Variance Filter, the High Correlation Filter, the Random Forests/ Ensemble Trees, the Backward Feature Elimination and the Forward Feature Construction, among others.

3.3.5 Feature Engineering

When working with sensory data, it is many times essential that feature engineering methods be applied, in order to bring the data in a tangible format and extract from them useful and measurable features. Methods like these are usually necessary in cases of signals, such as the Photoplethysmogram that presents the changing proportion of the light's absorbance during the analysis of the pulsed arterial blood with pulse oximeters. Engineering techniques could contribute to the extraction of features that would perform as assessment measures of the signal's quality. Therefore, in such cases were waveforms (vectors) are studied rather than a scaled size, observing measures like the mean or the variance would not be of any help. In signal processing it does worth studying the periodicity in the measurement's values. For this purpose Fourier Transformations could be applied and represent any sequence of measurements as a combination of sinusoid functions with different frequencies [2].

For the needs of the thesis project, a filter, the Savitzky-Golay one [43], was primarily applied on the continuous PPG signals to smoothen them and, as a result, repair the false small peaks of the signal. Smoothing the time series signifies replacing their values with new ones, obtained from a polynomial fit [44].Then, Fast Fourier Transformation was applied, in order to convolve the smoothened PPG signals and find the autocorrelation, a measure that signifies the similarities between the observations on the signal, as function of the time. The goal of the autocorrelation analysis is to provide findings of repeated patterns like the signal's periodicity, as performed in the present study.

3.3.6 Data Standardization

Many times, due to data heterogeneity, their standardization is needed in order, not only to be able to compare measurements that have different units, but also to increase the performance of most of the modelling and analytical algorithms. Performing such a technique, the values of the data are modified with a specific way to represent themselves in relation to the total and, in the end, reduce the data redundancy. More precisely, the data are rescaled to have a mean of 0 and a standard deviation of 1.

Standardization can be applied with many different methods like the standard deviation one, with which the mean is subtracted from all the data, and then they get divided with it resulting in reducing the variance of the sample, the Min-Max normalization where the data have to fit a predefined interval, the decimal scaling etc. [45].

In the context of this thesis project, the z-score normalization will be the method to be performed so that the data be standardized [46]. More specifically, z-score rescales the data to have the properties of a standard normal distribution, with mean (the average) $\mu = 0$, and standard deviation (from the mean) $\sigma = 1$. The z-scores can be calculated from the following equation:

$$z = \frac{x - \mu}{\sigma}$$

The "standard score" of z, if placed on a normal distribution curve, can indicate how many standard deviations below or above the plotted population's mean, an observation's score can be. Its values usually range from -3 to $+3$ standard deviations from the mean [47].

3.4 Exploratory Analysis

This part of the analysis can be performed with statistical and visualizing methods and aims in understanding the nature of the data and their principal characteristics by assessing assumptions for the analysis to be based and identifying the features – keys for the analytics procedure [48].

3.4.1 Descriptive Statistics

Examining the raw data, it is very difficult to identify specific patterns and visualize what they are presenting, so, in order to understand them, descriptive statistical methods need to be performed [48]. The huge variety of relevant techniques can be summarized in the following basic categories [49]:

- **Numerical Counts and Frequencies:** That indicate how many times an occurrence has been identified in the dataset.
- **Percentages:** That express a set of values as a proportion of the whole.
- **Measures of Central Tendency:** The mean, the mode and the median are the summary statistical single values that try to describe the dataset by identifying its central position.

  More specifically, the mean value of a set of n observations is defined as the sum of the total observations, divided by the number n of the total observations. Therefore, in a sample of size n, if the observations of the variable X are $x_1, x_2, \dots, x_n$, the mean value μ will be calculated as follows:

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Measures of Variability:** The range, the interquartile range, the standard deviation and the variance are the measures that attempt to describe the amount of variability or spread in the dataset.

  As far as the standard deviation σ is concerned, it constitutes the square root of the variance. The variance s^2 of a set of n observations $x_1, x_2, \dots, x_n$, is defined as the average of the squares of the deviations of the observations from their mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  Consequently, the standard deviation is calculated as follows:

$$\sigma = s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- **Measures of Skewness and Kurtosis:** Skewness constitutes the measure of symmetry of the probability density function in the distribution of the data and Kurtosis is the measure that describes if the data are heavy or light tailed, in relation to a normal distribution. Kurtosis performs also as an indicator of the possibility to have outliers in the dataset [50].

In the context of the thesis project, the mean and the standard deviation, that constitute the most valuable statistical measures, are used as indispensable inputs for many algorithms and techniques that are being performed. Of great importance in the context of the thesis are also the measures of Skewness and Kurtosis that are extracted as features during the processing of the PPG signals.

3.4.2 Inferential Statistics

The complex approach of inferential statistics gives information about significant differences in the set of data and reaches conclusions that could be extended from the examined data to the whole population. Some of the major methods that are used for the purposes of the statistical analysis, include:

- **Correlation:** This analytical method is used in order to determine the relations among the observations in the set of data and characterize them as negative positive, weak, or statistically significant [48]. In such problems, the strength of the correlation depends on the absolute value and the closer it is to 1 or -1 the stronger the correlation. As a result, this method indicates patterns that lead the analyst to control or perform forecast of future phenomena. Such relations could be the correlation of the variables, the correlative dependence, mutual restrictions etc. and in the context of the current thesis they are identified via heatmap and scatter plot visualizations.

- **Kolmogorov-Smirnov:** The non-parametric test that is used for testing if a variable follows a given distribution, i.e. the normal, in a population, or if two different populations come from the same distribution [50].
- **Welch's t-Test:** Constitutes a one-sided or two-sided, two-sample t-test that determines if the means of two populations are equal, in case that the variances are unequal or unknown. The test requires that the independent variable be categorical, the dependent variable be continuous and measured on an interval or ratio scale and that the distribution of both the two populations follows the normal distribution. The null Hypothesis is that the means of the two datasets are equal and they are rejected when the calculated p-value is less than 0.05 [51].

3.4.3 Visualizations

In order to be able to determine the causality of the data, make comparisons among the variables, understand them easily, perform different analytical tasks, or communicate the occurring results to the end user, data visualizations would be very helpful.

The graphical techniques that are used widely for the purposes of the thesis are the line bars for timeseries, ranking or deviation visualizations, the histograms for data or frequency distributions or skewness and kurtosis visualization, the scatter plots or the heatmaps for the illustration of correlations among the variables, the choropleth maps for geospatial presentations, the boxplots for outlier values detection, etc.

3.5 Machine Learning

Having conducted the descriptive statistics and having gained thorough comprehension of the data, it is time for the analysis. This part of the overall procedure can be performed not only with statistical methods, but also with the exploitation of data mining algorithms. The inferential statistical approach focuses in drawing conclusions from a data object, to its random variations, while the algorithmic one exploits the machine learning techniques that allow the extraction of unknown or hidden, possibly useful information and acquire knowledge from random, noisy and massive data.

As is has already been presented so far, the need to mine more and more data and extract information from them, is enormous. At the same time, artificial intelligence (hereafter, AI) has gained a wide scope in recent years, aiming, basically, in addressing the problem of over-information. This phenomenon has led scientists to develop systems, methods, techniques and algorithms that can automatically process and filter out the ever-increasing amount of data, giving birth to new knowledge that can be derived out of them. This evolution of research on pattern recognition techniques and the computational theory of learning in AI, has given rise to the field of machine learning, the field that allows the computers to “learn”.

More specifically, machine learning makes it possible to build adaptable computer programs running on the basis of automated analysis of data sets, based on the theory of statistics and data mining. The algorithms of machine learning that can handle most of the important problems in the data mining research, such as classification, clustering, regression, statistical learning, association analysis, linking mining, etc. can be distinguished in categories of learning depending on the desirable result, the most common of which are the supervised and the unsupervised one.

3.5.1 Supervised Learning

Supervised learning refers to the task of learning a function to represent the given inputs to known or desirable outputs (training set), with the ultimate goal of generalizing this function for inputs with unknown output (control set). In other words, all the data are labeled, and the algorithm uses the input data to predict the output. This approach of learning can also be evaluated and measured in terms of accuracy and efficiency, as the datasets that are used for the algorithm's training are known.

The problems that belong to this category of learning can be discretized in classification and regression ones.

- **Classification**: This task is used when the desirable output variable is a discrete label, a category. A classification model reaches conclusions from the observed values, and tries to predict categorical class labels. Some of the most characteristic classification models are the linear ones, like support vector machines (SVM) and the Naïve Bayes classifier, neural networks, decision trees, such as random forests, kernel estimation classifiers like the k – Nearest Neighbors (k-NN) etc.
- **Regression**: This task is used for the prediction of output variables that are continuous, real values, and it understands which independent variables are related to the dependent, exploring the relationship among them. In other words, regression analysis can determine if a variable can be used as a predictor for others, revealing the hidden dependence relationships among them [48]. Some of the most popular algorithms that handle such problems are the linear regression, the nonlinear regression, generalized linear models, etc.

3.5.1.1 k – Nearest Neighbors (k-NN)

In the context of the current project, the k-NN algorithm has been implemented with the scope of performing imputation of the mean. More specifically, the algorithm is based on the idea that the datapoints that are valued with similar characteristics have the tendency to belong to the same class. To predict this class, the algorithm examines the k-nearest datapoints (neighbors) in relation to the testing datapoint and predicts their most common class. The distance is determined by the position of the sample's datapoints in the n-dimensional space, where each dimension is assigned with a characteristic attribute.

The proximity to the neighbors can be calculated with measures such as the Euclidean distance for continuous variables, the Manhatan or the Mahalanobis ones for discrete variables, and thus, the algorithm can predict both discrete and continuous attributes, either by selecting the most frequent value among the k – neighbors, or by calculating the mean from them [52]. The most crucial issue regarding this technique is to choose the appropriate k [53] and, for this purpose, there are several possible approaches:

- The missing value can be replaced by the value of the 1st closest neighbor, having a k equal to 1.
- The k can be equal to the total number of datapoints in the sample. Although this approach reduces the risk of overfitting by suppressing the noise effects, it makes the classification boundaries barely distinct.
- In many cases the k is calculated as the square route of the total number of datapoints in the sample
- In most cases and odd value is selected as k so that confusion be avoided between the classes.

The main advantage of the method is that it can easily approach the target function and that it can be even more easily programmed. Also, if small changes take place in the training, sample set, the classification result does not get particularly affected.

3.5.2 Unsupervised Learning

In this machine learning process there is no predefined set of values. The training set is divided into groups that are initially unknown, with unique criterion their key – characteristics. The algorithms learn to inherent the structure from the input data and approximate a function that describes the hidden structured of the unlabeled data. Since the samples that are used in this learning model are not labeled, there is no error and, hence, no evaluation mark to assess the structure that the model finally discovers as a possible solution to the problem. Besides, this constitutes the most significant difference between the two types of learning.

The vast majority of unsupervised learning problems can be found in clustering analysis, among others.

Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping the patients by their adherence in the measurements' schedule. Cluster analysis is used for exploratory data analysis to find hidden patterns or groupings in data. The clusters are modeled using a measure of similarity which is defined upon metrics, like distances. Since the notion of "cluster" is difficult to be defined precisely, various algorithms have been created in order to satisfy this need and, depending on the cluster model and its particular properties, they can be divided into general categories.

Among the various different clustering models, some of the most common are the connectivity ones, such as the hierarchical clustering, where the models are built based on the distance connectivity, like Euclidean, Manhattan or Mahalanobis, the centroid models, like the k-means algorithm, where the clusters are represented by vectors, and the density models, like the Density-based spatial clustering of applications with noise (DBSCAN), that defines the clusters as connected dense regions in the space of data.

The needs of the analysis that is implemented in the context of the thesis project, detect the use of the unsupervised clustering method to be performed, as it attempts to find a kind of structure, in terms of clusters, in the unknown data.

3.5.2.1 Clustering with k-Means

This clustering method aims at partitioning a given a set of n observations $X = x_1, x_2, \dots, x_n$, where each observation is a d – dimensional real vector, into k disjoint clusters $C = C_1, C_2, \dots, C_k$. As a result, each observation belongs to the cluster with the nearest mean μ_j, called centroid, and serves as a prototype for the cluster. In order to find the optimal prototypes, the within-cluster sum of squares, also known as inertia, should be minimized. And choosing the appropriate centroids that minimize the inertia, the measure of internal coherence between the clusters, is k-Means' biggest challenge [54]. The equation that performs this need is the following:

$$\sum_{i=1}^{n} \min_{\mu_j \in C} \| x_i - \mu_j \|^2$$

The later formula intends to find clusters that are different with each other and every observation in each cluster should be similar with the ones that belong to the same cluster.

3.5.2.2 Hierarchical Clustering

Hierarchical cluster analysis is another unsupervised technique used to cluster unlabeled data points and it is based on their similar characteristics. Depending on the clustering approach of the data, the method can be categorized into Agglomerative and Divisive. The 1st approach is bottom-up and begins from the particular data points. The 2nd one is top-down and it treats the data points as a big cluster that through the clustering process is being divided into smaller ones [54].

Within the framework of this project, the Agglomerative approach is the one that has been implemented. Performing such a data clustering technique, the hierarchy of clusters is represented as a tree, a dendrogram, where each leaf corresponds to one object, and, moving bottom-up, one can observe the groupings of similar objects into larger clusters, the branches, which – based on their similarity – are bonded to each other while moving to a higher level, and, in the end the procedure results in having all the initial data points included in a single, big cluster [55]. The fusion's height indicates the dis-similarity between two objects and the higher the height of the fusion is, the less similar the objects are.

The most successful method to perform an agglomerative hierarchy cluster analysis is the complete linkage, also known as farthest neighbor clustering [56]. At the beginning, each observation constitutes a unique cluster. Then, all the clusters are sequentially combined into bigger ones, based on the shortest distance between the observations in each pair of them. The distance is defined as the link between the two observations, one from each cluster, that are farthest away from each other. The shortest link is the one that fuses the two clusters, the observations of which are involved. This procedure is repeated until all the observations belong to the same cluster [57]. The mathematical equation that describes the distance between the observations of the clusters, is the following [58]:

$$d(u,v) = \max(dist(u[i], v[j]))$$

for $i \epsilon u$, $j \epsilon v$ and u, v the two clusters.

4. CASE STUDY

Taking into consideration the literature review about the non-communicable diseases, it has become apparent that the necessity to monitor a chronic patient’s health condition, constantly, is of imperative need. It has also been obvious that the exploitation of the Quantified Self – trend of our times, the ever increasing “Big Data” and the ever-evolving new technologies of IoT is of immediate priority, as it can play a decisive and vital role in disease control. Within this framework, the Greek startup company “BioAssist S.A.” implemented in 2014 an online health service, called “HeartAround” [59].

4.1 The “HeartAround” Service

The “HeartAround” service is addressed to elderly patients that suffer from severe chronic diseases. With the aim of offering the benefits of mobile health to the patients, it also provides them the potential of self-managing their medical condition. The implemented service enables the recording and processing of vital signs of patients suffering mostly from high blood pressure, pulmonary diseases, diabetes and cardiovascular diseases and, at the same time, controls their physical activity and captures their emotional status, among other functionalities.

4.2 Technologies and Devices

In the context of enhancing the independent living of the involved users who live in remote areas of Greece, BioAssist incorporated various wireless and wearable sensor devices in a tablet application, in order to help them record the measurements of their vital signs in real-time, upon doctor’s instructions.

All the data are automatically stored wirelessly in a cloud infrastructure. Thus, the patients’ personal electronic health record (hereafter, PHR) is being created, a fact that can assist in their appropriate personalized medical follow-up and treatment. The cloud platform is based on the popular framework Node.js and stores the PHR in a NoSQL database, MongoDB. The service exploits several Bluetooth wireless technology devices like pulse oximeters, blood pressure monitors, electronic scales, spirometers, glucometers and wearable activity trackers, in combination with an Android tablet device.

4.3 Pilot Programs

BioAssist, in the context of verifying the feasibility of the implemented service, initiated several pilot programs in cooperation with the relevant departments of University Hospitals all over Greece that, over a period of time, became permanent. Table 3 presents these pilot programs, categorized into groups, based on the geographical area where they took place, along with the pilots’ start date and the chronic diseases that the involved users suffer from.

Table 3: Pilot Programs

Patient Programs	# of Patients	Diseases	First Day of the Program
Pilot Athens “BioAssist Friends”	34	- Healthy - Chronic Obstructive Pulmonary Disease (COPD) - Hypertension - Diabetes	25/11/2014

Pilot Larisa	34	- Idiopathic Pulmonary Fibrosis (IPF) - Pulmonary Embolism - Emphysema	19/01/2016
Pilot Crete	8	- IPF - Emphysema	03/01/2017
Pilot Thessaloniki	7	- IPF	30/11/2017

The enrolled users are provided with an Android tablet with the “HeartAround” application pre-installed, a set of devices to record their biosignals, depending on their disease, and, in some cases, a smartwatch. Each user has a pre-defined time schedule of the measurements they need to take, as indicated by their doctor’s instructions, that needs to be repeated on a daily basis.

5. METHODOLOGY, EXPERIMENTS & RESULTS

Taking into consideration the scientific review and based on the technical background of data analytics, this chapter will present the methodology that has been followed in the context of the current thesis project, the experiments that have been performed and the obtained results. A general overview of the procedure can be reflected in the following figure:

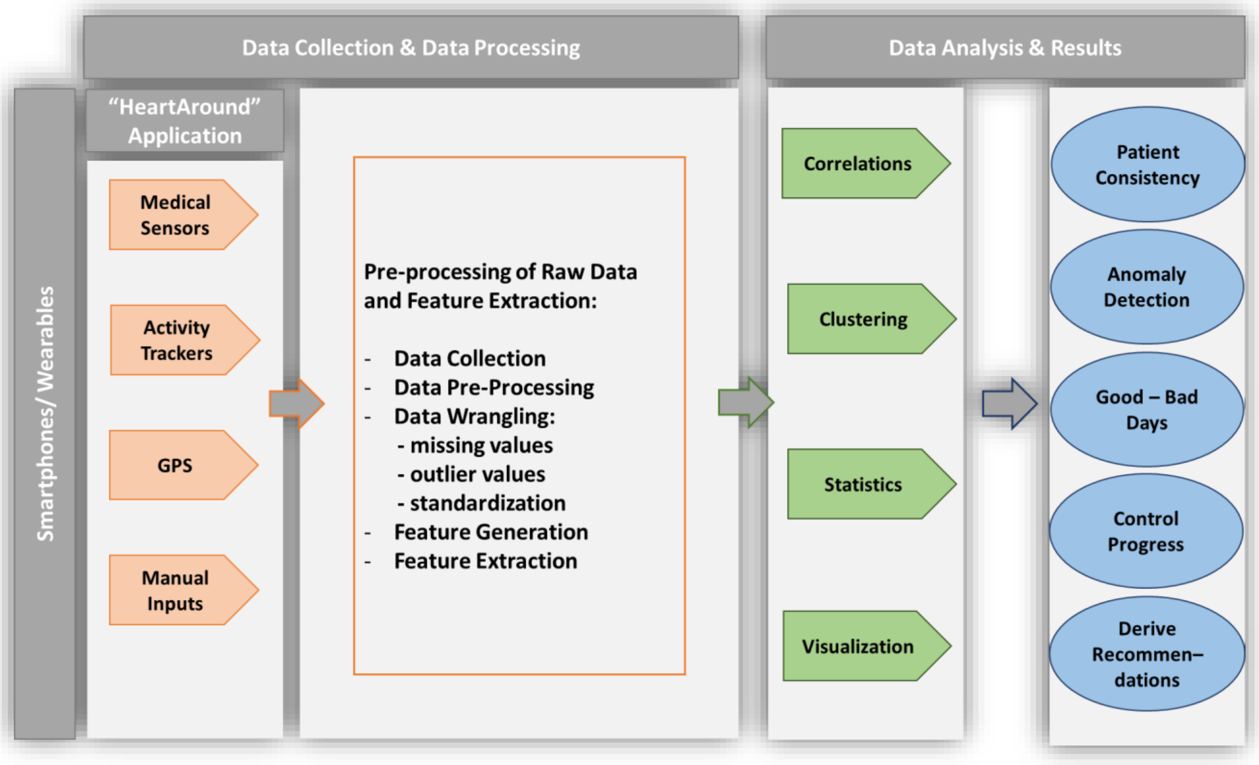


Figure 10: Thesis Methodology

5.1 Scope of Analysis

Within the framework of the thesis project, it has become apparent that the analysis of the data related to a patient's health condition is of great importance in order to understand them and extract the maximum possible information from them, with the scope of answering some basic questions regarding:

- The types of data that occur and the population that these data occur from, in order to be able to locate our findings (i.e. sex, age and geolocation),
- The interaction among the different measurement categories and if this interaction can be correlated with the patients' medical situation, in order to identify possible risk factors and, as a result, be able to derive relevant recommendations,
- The frequency of the measurement – taking, in order to determine if the patients' consistency or inconsistency can have repercussions on their health condition, and
- The quality of the measurements in an everyday basis, in order to enable the patients self-monitor the progress of their health and early detect dangerous situations.

With the goal of investigating the above case studies, we used various techniques, methods on an appropriate dataset and conducted an in-depth data analysis.

5.2 Data Collection

Starting off with the fundamental material, the data, BioAssist kindly provided us with a sub-dataset of their database, appropriate enough to fulfill the purposes of the analysis performed in the context of the thesis project.

The dataset will be presented as follows:

5.2.1 BioAssist's Dataset

The collected data, stored in a CSV file, involve different attributes related to the different patients who use the "HeartAround" service. Each of these attributes have been already presented in the paragraph 2.1.3 and for every one of them there is a measurement that has been recorded at a specific timestamp. Most of them are single values that vary in type, from numerical to categorical. These attributes are going to be analyzed thoroughly through the procedure of pre-processing, in the upcoming section.

The dataset combines the needs of the current research project in studying chronic diseases and quantified self – related information, and the types of data it includes can be summarized in Table 4:

Table 4: Data Studied in the Current Research Approach

Area of Interest	Category	Data Recording	Attribute	Type
Healthcare	Vital Signs	Pulse Oximeter	SpO2	Numerical
			PPG	Numerical (Signal)
			Heartbeat	Numerical
			Perfusion Index	Numerical
		Blood Pressure Monitor	Systolic BP	Numerical
			Diastolic BP	Numerical
		Glucometer	Blood Glucose	Numerical
		Spirometer	FEV1	Numerical
			FEV1_FVC	Numerical
			FEV6	Numerical
			FEF2575	Numerical
			FVC	Numerical
			PEF	Numerical
		Scale	Weight	Numerical
Fitness	Activity	Activity Tracker	Steps	Numerical

			Sleep	Numerical
Mood	Feeling	Manually	Current Emotional Status	Categorical
Demographics	Personal Data	Tablet Service with GPS	Year of Birth	Numerical
			Sex	Categorical
	Location		Zip Code	Numerical

5.3 Data Analysis

As it has already been mentioned in the previous chapter, the “HeartAround” service is being used by elderly people who live in different areas of Greece and suffer from various diseases. What is more, since one of the service’s goals is to enhance the independent living of the involved patients, they are urged to use the service’s components, the tablet and the wireless sensors, on their own and remotely, from their home. Also, each patient follows their personalized measuring schedule, which is, probably, not the same for all the different types of measurements they take. Last but not least, it is undeniable that, due to users’ negligence or due to devices’ malfunctions, there will be days with no measurements at all. As a consequence, it is anticipated that the dataset be heterogenous, with missing data, outlier values, multiple recordings during the same timestamp, and so on. In order to explore the dataset and reach safe conclusions about the patient’s condition, we needed to move from the raw data that BioAssist collected and provided to us, to a cleaned dataset, adequate to be used in statistical and machine learning analytical techniques.

5.3.1 Data Overview

Having a glance at the dataset, we observed that it consists of 16 different columns, the information of which can be summarized in Table 5:

Table 5: Columns of the Initial Dataset

ID	Unique identifier per record
MASTER_ID	Unique identifier of the biosensors that take multiple measurements of different vital signs
PATIENT_CODE	Unique identifier per user
DATE	The date that each measurement took place
TIME	The time that each measurement took place
DEVICE	The medical biosensors or activity trackers that have been used to record the measurements
READING_VALUE **READING_VALUE2** **PPG_ARRAY**	Values of the different biosignals’ measurements
VALUE_UNIT **VALUE2_UNIT**	Unit of measurement

GENDER	1 stands for male 2 stands for female
ZIP_CODE	The zip code of the geographical area that each person lives
CATEGORY	The group that each patient belongs to

Delving deeply into the dataset, we found out that it includes measurements of 168 people that are not only patients, belonging to different groups related to: the location they live, the disease they suffer from and, in case they are not patients, the role they have in the company (i.e. testers of the “HeartAround” application’s functionality). We also identified hundreds of missing values and noticeable time inconsistencies, among the various measurements. Table 6 can give a brief overview of our findings:

Table 6: Overview of the Initial Dataset

Date of 1st Measurement	14/ 11/ 2014
Date of last Measurement	17/ 12/ 2018
No. of Involved Users	**168** people
No. of Groups	**16** different groups
No. of Measurement Categories	**17** different reading categories
Total No. of Records per Day	**956.204** records

5.3.2 Pre-Processing

So far, we have had a first impression that our initial suspicions were most likely to be proved correct, since the dataset was even more heterogenous and complex than we expected. Thus, in order to put the data into an order and have a better image of the patients and the information that regards them, we performed several modifications on the dataset, guided by the results we were given each time.

More specifically, we started by fixing the “NULL” values in the “READING_CATEGORY”, based on the “VALUE_UNIT” or the “DEVICE” that correspond to the same record. Then, we replaced the nulls in the “CATEGORY”, “GENDER” and “YEAR_OF_BIRTH” fields with the “unknown” value, we calculated the age of each participant and, last, we merged the column “PPG_ARRAY” (which has all the values of the “PPG” measurements) with the one that has the rest of the reading values.

As a next step, we split the dataset into three different subsets and we plotted the number of reading values performed from all the patients, per day. The first dataframe represents the initial dataset, the second one includes only the pilot programs of remote areas in which BioAssist participates (hereafter, “Remote Pilots”) and the third one represents the group “BioAssist Friends”. The later consists of both patients and healthy people, not supervised by a physician, constituting the first pilot users and performing as regular participants, regardless of whether they suffer from a disease or not (hereafter, “Pilot Athens”).

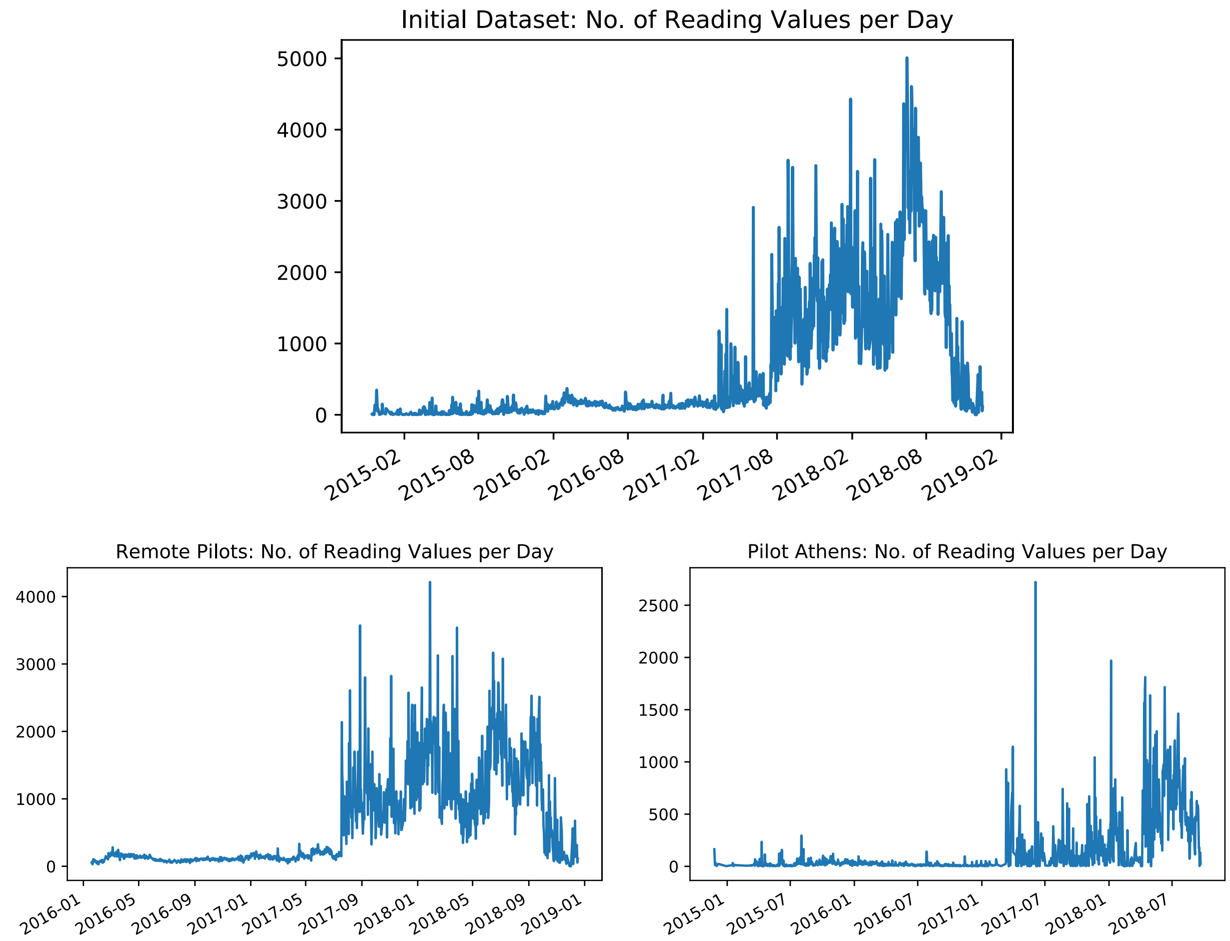
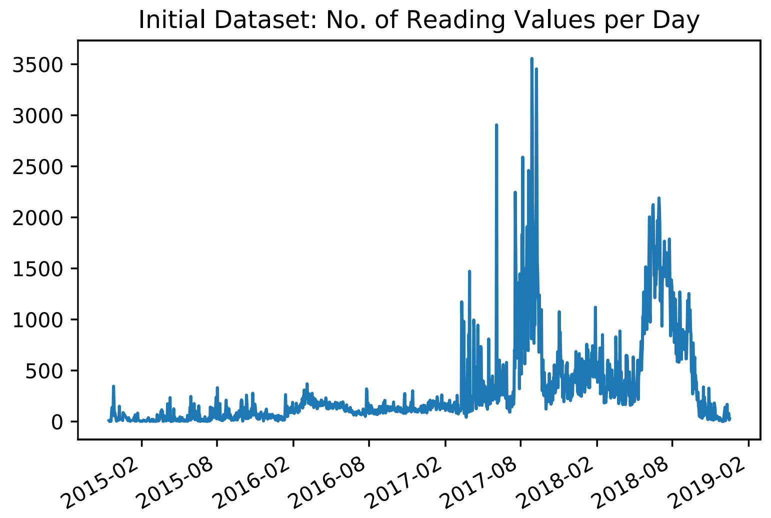


Figure 11: No. of Reading Values per Day

Figure 11 gave an indication of what is going on with the groups of the dataset and which is the leading one. In order to be able to prove it, we performed some further changes. Firstly, we merged the “READING_VALUE” and “READING_VALUE2” columns into one, we removed the needles columns, such as ‘ID’, 'MASTER_ID', 'PPG_ARRAY', 'READING_VALUE2', 'DEVICE', 'VALUE2_UNIT', 'ZIP_CODE', 'YEAR_OF_BIRTH', 'VALUE_UNIT', we pivoted the table of the data so that each reading category be in a separate column, we renamed some attributes’ labels so that they be more comprehensive and, finally, we grouped the data per date and time. Plotting the datasets as we did before, we got the following results:



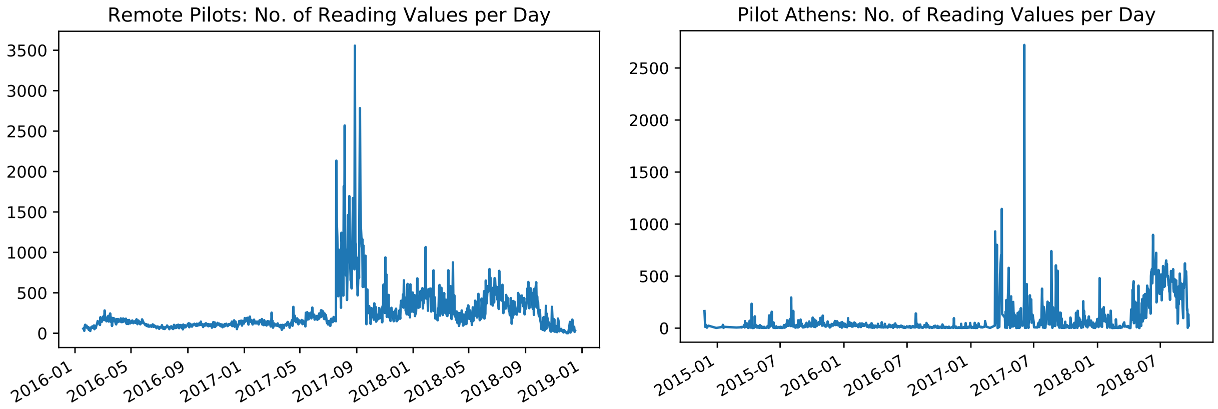
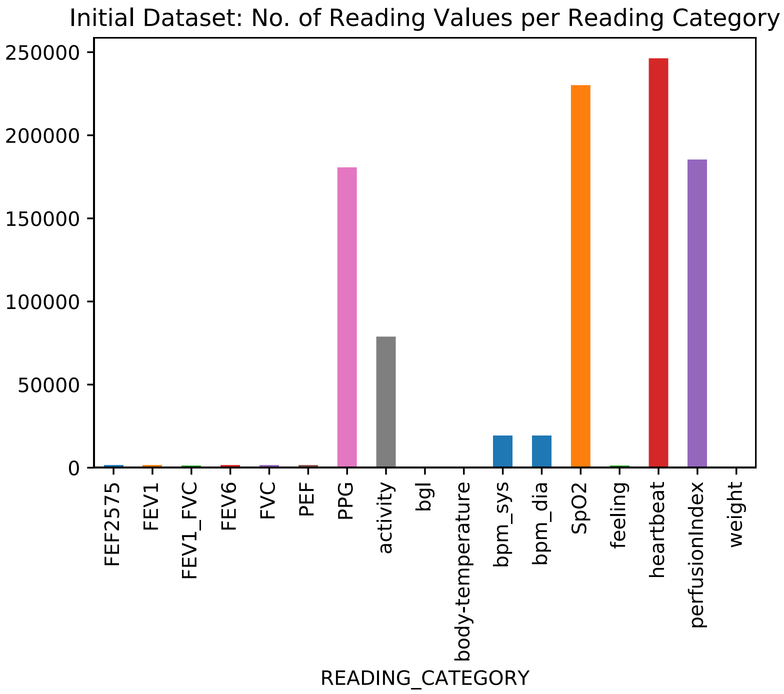


Figure 12: No. Reading Values Per Day, Grouped by DateTime

Observing the plots and performing some calculations, we understood that, even though the measurements of the pilot groups start 14 months after the first time the service was set to function, the Remote Pilots' 49 patients – out of 168 in total – have performed the 70% of the total readings of the initial dataset and the Pilot Athens' 36 patients, the 17%. The rest of the groups have so few measurements that it is needles mentioning them. Undoubtedly, the pilot groups where the ones that we would keep in order to continue with the analysis.

Moving on, we wanted to check which of the 17, in total, reading categories has the majority of records in the dataset. Using a bar plot this time, we visualized the number of total readings per reading category and we got the following results:



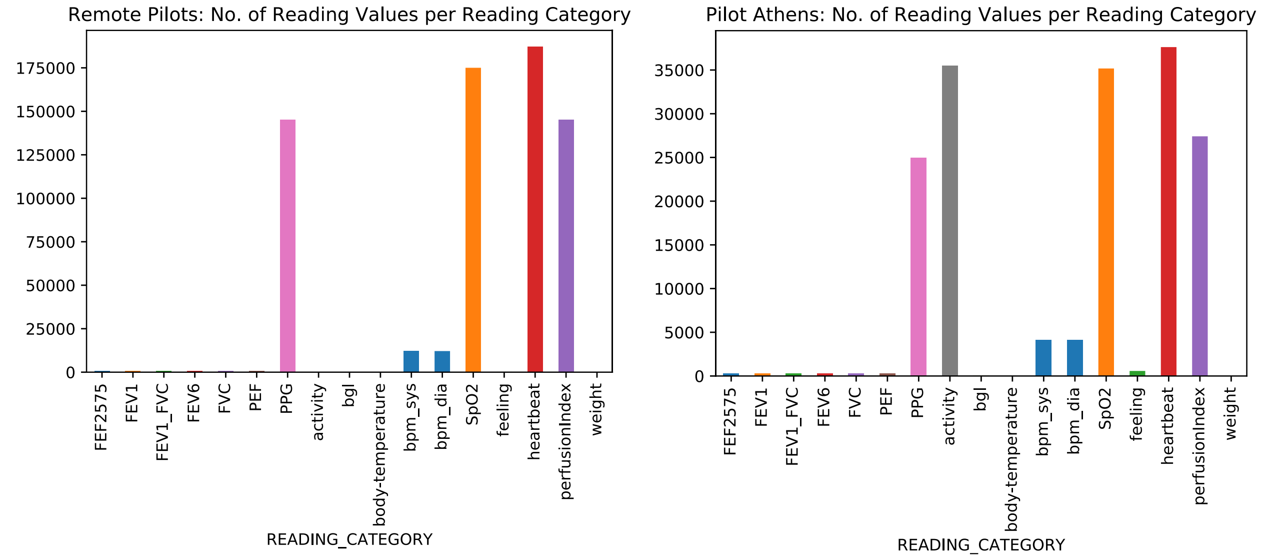


Figure 13: No. of Total Readings per Reading Category

Studying Figure 13, we can see that the reading categories that have the majority of the performed measurements, in descending order, are: Heartbeat, SpO2, PerfusionIndex, PPG, Activity, Bpm_sys and Bpm_dia. Also, the 6 reading categories that occur from the spirometry, the FEV1, FVC, FEV1/FVC, FEF2575, FEV6 and PEF should note be neglected, as they are very important indicator factors in cases of patients suffering from pulmonary diseases, like the ones from the remote pilots, and since this is not a daily measurement, it is normal that they don’t have so many records. Hence, we decided that these reading categories should be part of the analysis.

Last but not least, we made some further modifications to the dataset, such as appending the consecutive readings of the PPG signal (as explained in section 6.5.1) and extracting from the united PPG signals the features of skewness, kurtosis and their period, so that we could be able to include them in the analysis. We also dropped the reading categories: 'feeling', 'body-temperature', 'bgl' and 'weight' because they have too few records, along with the 'activity' one, since the values’ heterogeneity and complexity makes them incomprehensible, therefore, inappropriate for our study. Lastly, we kept the categories that we mentioned before and, thus, we created a new dataset, ready to be used for the analysis.

The following table shows the results of the data pre-processing step:

Table 7: Before – After Data Pre-Processing

	Before Data Pre-Processing	**After Data Pre-Processing**
Date of 1st Measurement	14/ 11/ 2014	25/ 11/ 2014
Date of last Measurement	17/ 12/ 2018	17/ 12/ 2018
Involved Users	**168** people	**85** people
Groups	**16** groups	**9** groups
Measurement Categories	**17** reading categories	**16** reading categories
Total Records	**956.204** records	**375.189** records
Average Age	**-**	**58.65** years old

5.3.3 Distinguish Consistent Patients

Making a quick review on the dataset that occurred through the procedure of pre-processing, it consists of eighty five people – most of which are patients that follow a doctor's guidelines (hereafter we will use the term "patients" for all of the pilots' participants) – that belong to nine different groups, constituting the four pilot programs in which BioAssist is enrolled, their mean age is 59 years old and during the period of four consecutive years, they have performed a total of approximately 375.000 measurements of thirteen different vital signs or activity – related reading categories.

5.3.3.1 Data Overview

At this point, we needed to understand further what is going on with each one of the patients. Thus, we printed the number of total readings of each patient, grouped by "READING_CATEGORY", and we plotted them per day, grouped by "READING_VALUE".

The results showed that among the eighty-five patients, some of them present consistency in keeping a schedule for their measurement taking, others take measurements randomly, once in a while, and some others have only one, or 20 measurements in total. Indicatively, we present some of the most representative examples:

Patient 6 from Pilot Athens has performed only one single measurement during the last four years.

FEF2575	0	PPG	0
FEV1	0	bpm_sys	0
FEV1_FVC	0	bpm_dia	0
FEV6	0	SpO2	0
FVC	0	heartbeat	1
PEF	0	perfusionIndex	0

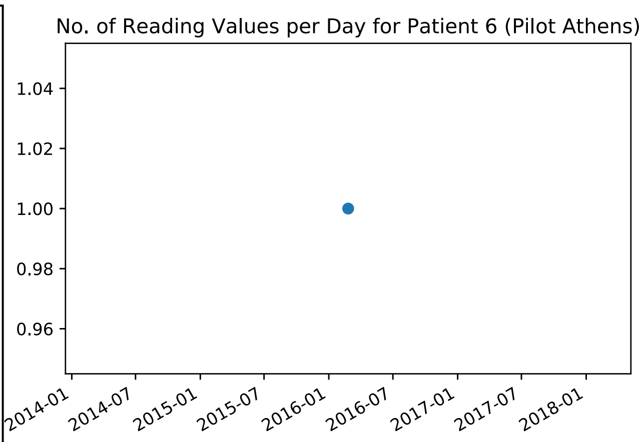


Figure 14: Patient 6 – Reading Values per Day

Patient 17 from Pilot Athens. Although a lot of measurements have been performed since the day that the patient started being part of the program (~3 years), the plot indicates the patient's inconsistency in keeping a measurement – taking schedule.

FEF2575	19	PPG	736
FEV1	19	bpm_sys	747
FEV1_FVC	19	bpm_dia	747
FEV6	19	SpO2	1551
FVC	19	heartbeat	1699
PEF	19	perfusionIndex	736

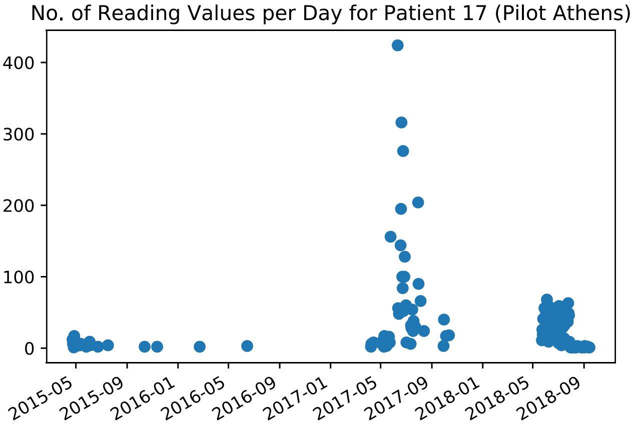


Figure 15: Patient 17 – Reading Values per Day

Patient 47 from Pilot Crete has performed the following measurements since the day they started being part of the program (~1 year).

FEF2575	553	PPG	23450
FEV1	553	bpm_sys	287
FEV1_FVC	555	bpm_dia	287
FEV6	553	SpO2	23449
FVC	553	heartbeat	23734
PEF	553	perfusionIndex	23450

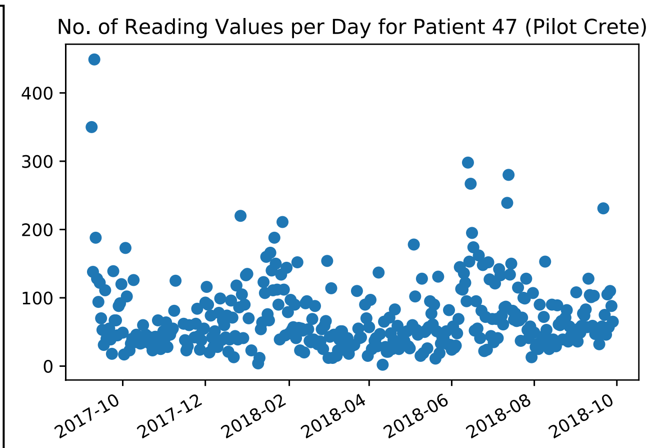


Figure 16: Patient 47 – Reading Values per Day

Patient 3 from Pilot Larisa has performed the following measurements since the day they started being part of the program (~2.5 years).

FEF2575	8	PPG	3996
FEV1	8	bpm_sys	926
FEV1_FVC	6	bpm_dia	926
FEV6	8	SpO2	5322
FVC	8	heartbeat	6247
PEF	8	perfusionIndex	3996

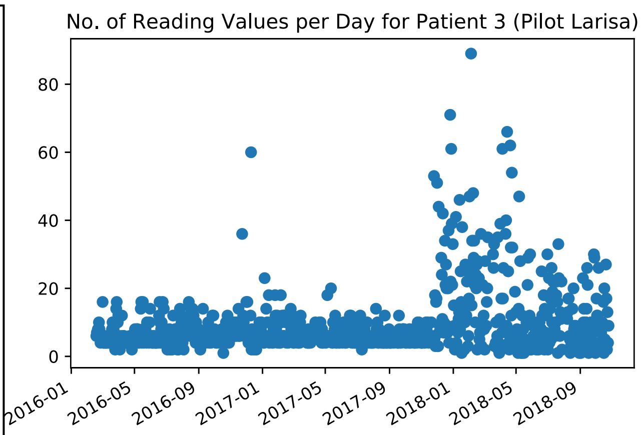


Figure 17: Patient 3 – Reading Values per Day

An overall picture of the eighty-five patients' performance in measurement – taking can be reflected in Figure 18. Observing the plots that correspond to each patient, it becomes evident that we have a very heterogenous dataset, and this is the most crucial point: to choose the right patients, whose records will help the research reach safe conclusions. Consequently, we needed to distinguish and separate the consistent from the inconsistent patients and only then to proceed with the former to the next steps of the analysis. For this purpose, an unsupervised clustering technique was essential to be used.

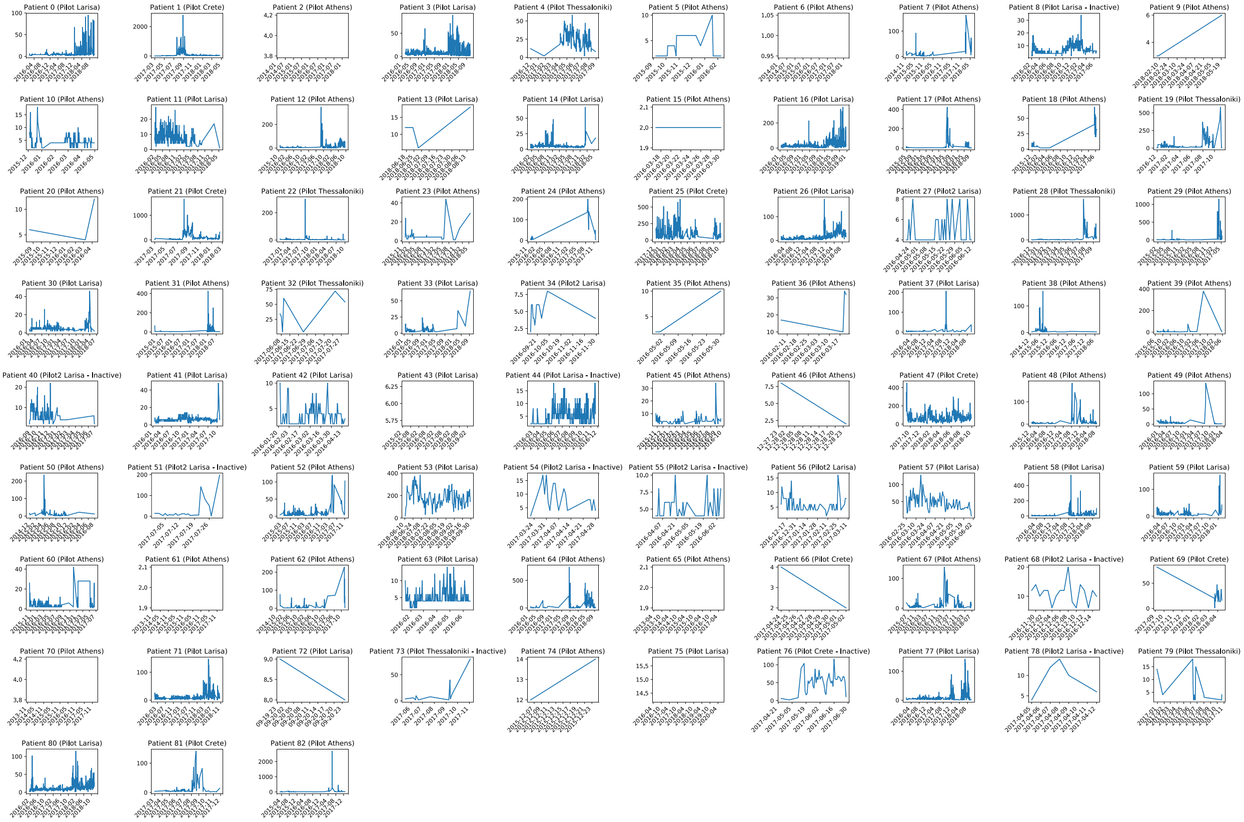


Figure 18: Reading Values per Day for the 85 Patients

5.3.3.2 Feature Extraction

To begin with the clustering procedure, we needed to extract the most representative features of the dataset. After thorough study of the findings we had so far, we decided to calculate, produce and use the following four features to be used for the clustering:

Table 8: Extracted Features for Clustering

Feature 1	Mean of Day Differences	The mean value of the days between two days that the patient took a measurement. This answers the question: “How many days has it been since the last time the patient took a measurement?” and, obviously, is a measure to be compared among the patients and indicate their frequency.
Feature 2	Total Days	The number of days that each patient has taken measurements of their biosignals. Since we saw that there are patients that have taken very few measurements in only 5 or 10 days in total, this measure would be very helpful.
Feature 3	Total Readings	The number of measurements that each patient has taken in total. Based on a logical assumption, if a patient takes a lot of measurements, it is highly possible that they are consistent in keeping a program of measurement – taking.

Feature 4	Total Days/ Total Readings	The ratio of total number of days per total number of readings. If the ratio yields a small value, this means that, compared to the total number of days, we have a lot of measurements, which would be an indicative factor regarding the patient's consistency.

From the above features, the one that constitutes the strongest chip is the first, as, even if a patient has taken a lot of measurements but not on a regular basis i.e. every 5 or 10 days on average, they cannot be included in the category of the consistent ones.

5.3.3.3 Handling Outlier Values

Since the clustering algorithms are very sensitive in outliers, we needed to detect them and exclude them from the dataset and, thus, eliminate the possibility of the procedure's failure.

For this purpose we performed for the features 1 and 4 the standard score (z-score) that normalizes the data and checks which values exceed the mean of the data distribution. If the mean of day differences is too high, 100 for instance, this would mean that the patient took measures every 100 days, on average. For sure, such a patient would not be consistent so we would have to exclude them in order not to influence the clustering results. Similarly, if the ratio is too big, this would mean that in a big period of time, the patient has taken only a few measurements, hence, they are not consistent and should be removed before performing the clustering algorithm. The results we were given are the following:

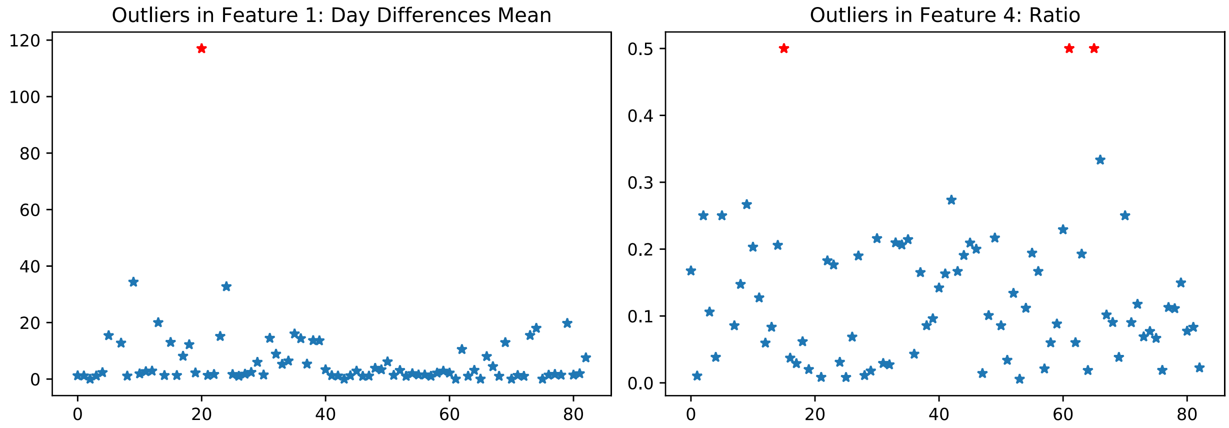


Figure 19: Detecting Outlier Values

Based on the visual results and the assumption that the normal values should be in the range of $-3 \leq z - Score \leq +3$, we removed the outlier values that exceed the allowed limits and, hereafter, we were able to procced with the clustering procedure.

5.3.3.4 Clustering

For this step of the analysis, we decided to perform two different clustering approaches, so that the one could verify the results of the other, to distinguish correctly the consistent from the inconsistent patients and, as a result, have the correct sub-dataset to proceed with, during this analytics project.

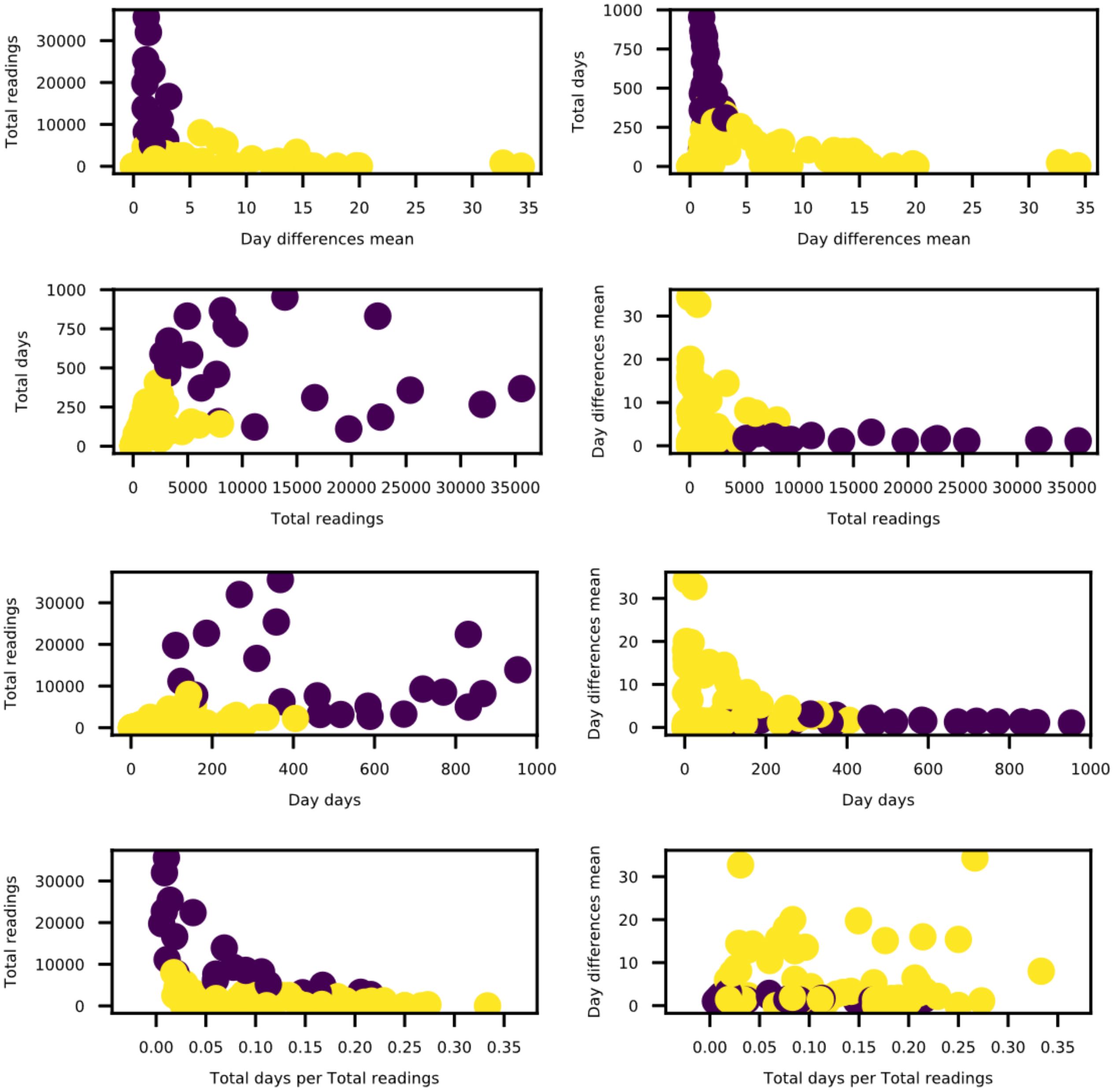
5.3.3.4.1 Standardization

The first algorithm that we implemented is k-Means. Since the variables of the dataset are of incomparable measure units, it was essential that the data be standardized before the implementation of the algorithm. The scaling of features has a great impact on measures such as the Euclidean distance that is used by k-Means. Since we need our features to have equal weights during the clustering procedure, we standardized them by subtracting their mean and dividing it by their standard deviation [60].

5.3.3.4.2 Clustering with k-Means

Setting up a k-Means with two clusters and applying the algorithm, we got 21 patients grouped into the 1st cluster, constituting the consistent patients and 37 into the second one, constituting the inconsistent ones.

Indicatively, plotting the clusters that we obtained for all the four extracted features, combined, each time, in pairs of two, we can see that the data are separated in two clusters: the one in purple color shows the consistent patients and the other one in yellow represents the inconsistent patients, as shown in Figure 20.



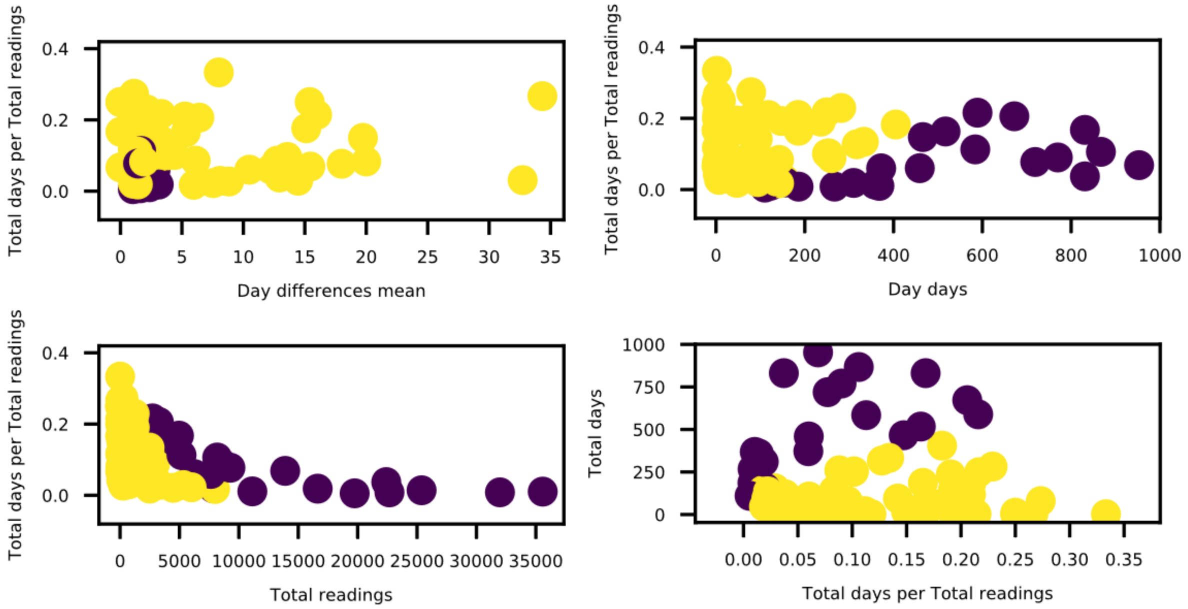


Figure 20: Consistency Results – Clustering with k-Means

5.3.3.4.3 Hierarchical Clustering

Having a first assessment on the patients' consistency issue, we continued by performing Hierarchical Clustering in order to verify the results. Having applied the algorithm to our data, we plotted the results we were given in the following dendrogram:

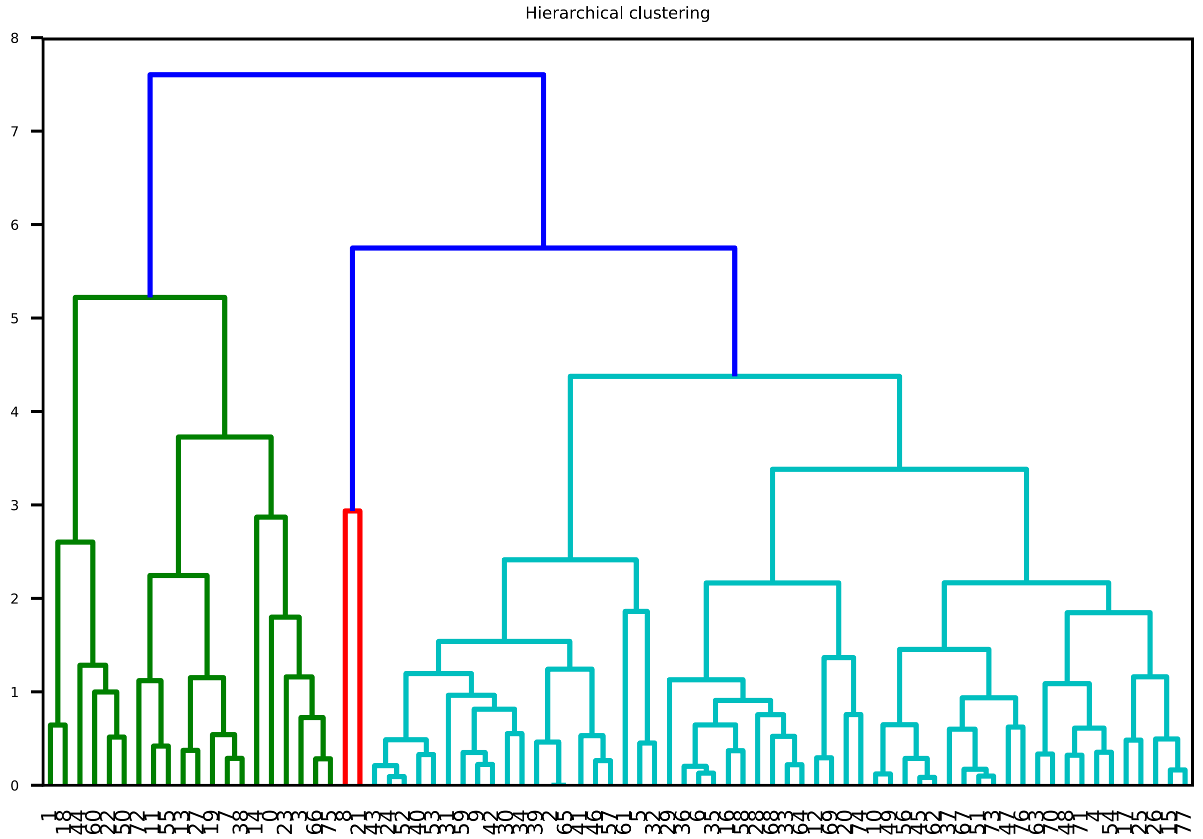


Figure 21: Consistency Results – Hierarchical Clustering

In the dendrogram displayed above, each leaf corresponds to one patient and moving bottom-up the tree, we can see the groupings of similar patients into branches (in terms of the features we are studying) which are bonded to each other as we are moving to a higher level. The fusion's height indicates the dissimilarity between the patients/ clusters and the higher the height of the fusion is, the less similar the patients are.

As a next step, we set the number of clusters that we wanted to divide the results of the hierarchical clustering and we got two groups containing 20 and 38 patients respectively.

5.3.3.4.4 Results Comparison

Comparing the results of the two different clustering methods, we observed that they present a slight difference, as they cluster only 3 specific patients to the opposite clusters.

Table 9: Clustering Results Comparison

Clustering Method	Consistent Patients	Inconsistent Patients
k-Means	21	37
Hierarchical	20	38

In order to decide which method has given the most realistic and representative results, we had to go back to the initial diagrams we drew during the data overview and study the plots of these three patients. Observing the features in the plots, in comparison to the algorithms' results, it seems that k-Means has given more accurate outputs that hierarchical clustering and, thus, we kept the k-Means clustering and its results to proceed with the analysis.

5.3.3.5 Dimensionality Reduction for Visualization

Having decided that k-Means is the best, against hierarchical clustering, method for data clustering, as far as our dataset is concerned, we wanted to validate the clustering performance, therefore we performed another visual inspection of the data. Since the characteristic features of the data we have extracted and studied are four, we could not visualize the results in one main plot, so, we applied the t-SNE algorithm that reduces the dimensionality down to two dimensions. Giving to t-SNE as input the output of the k-Means clustering, we were able to observe and validate that the data are separated well enough into two non-overlapping clusters:

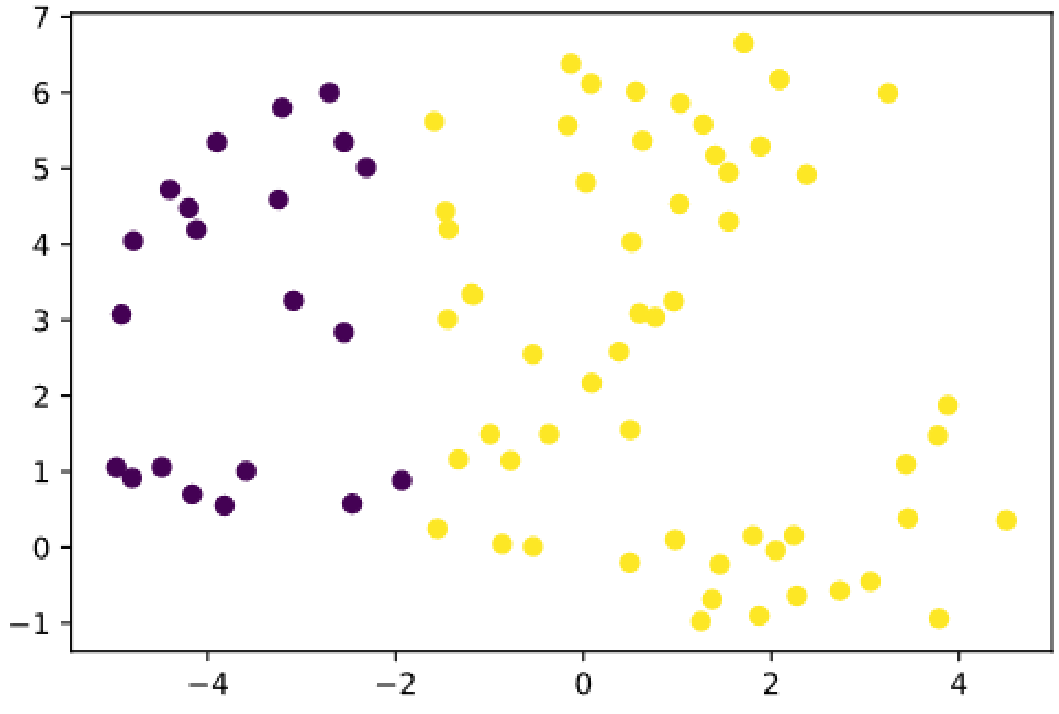


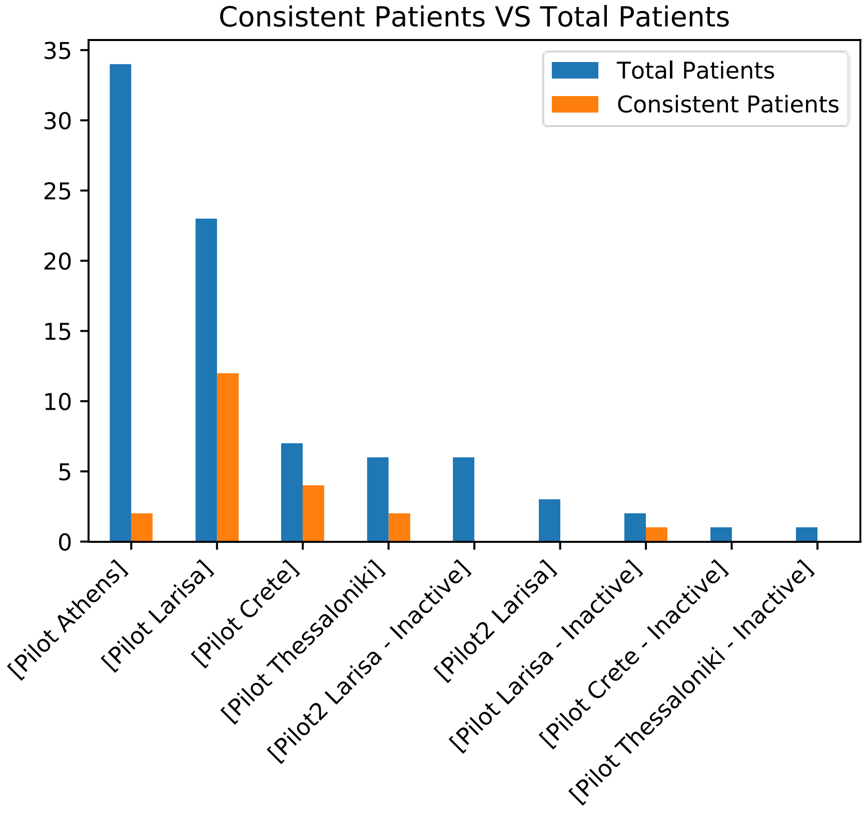
Figure 22: Consistency Results – Visualization with t-SNE

As a result, we kept the 21 patients that are proven to be consistent in keeping a schedule regarding their measurements – taking, and we saved them to α new dataset that will be used later. The following table summarizes the evolution during the analysis procedure:

Table 10: Before – After Data Clustering

	Before Data Pre-Processing	After Data Pre-Processing	After Data Clustering
Date of 1st Measurement	14/ 11/ 2014	25/ 11/ 2014	24/ 10/ 2015
Date of last Measurement	17/ 12/ 2018	17/ 12/ 2018	17/ 12/ 2018
Involved Users	**168** people	**85** people	**21** people
Groups	**16** groups	**9** groups	**5** groups
Measurement Categories	**17** reading categories	**16** reading categories	**16** reading categories
Total Records	**956.204** records	**375.189** records	**269.640** records
Average Age	**-**	**58.65** years old	**66.82** years old

What is worth mentioning is that from the 83, in total, patients that contribute to pilot programs, only 21 of them are consistent. However, these 21 patients have performed the 79.61 % of the total measurements. These patients belong to the pilot groups of Athens, Larisa, Crete, Thessaloniki and Larisa – Inactive, the proportion between men and women is 6/1, and 15 out of the 21 patients are older than 65 years old, as presented in the following figures:



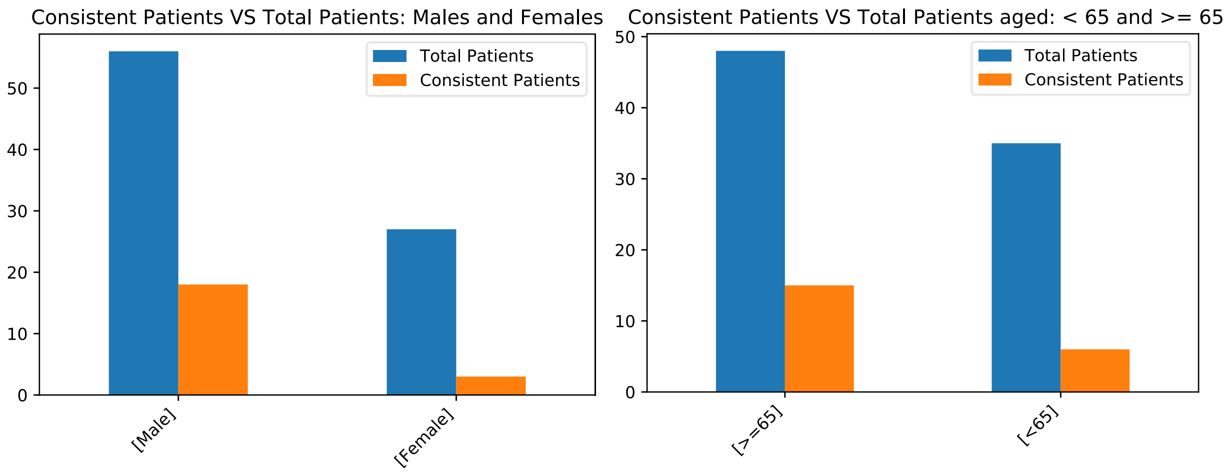


Figure 23: Consistent Patients VS Total Patients

5.3.3.6 Results Communication

Bearing in mind that these conclusions have been drawn by the analysis of the data that occurred from the use of an electronic health service, we wanted to provide a visualization that would fit this context and could be used as a web interface by the end user. Consequently, Figure 24 presents the geographic location of the pilot programs and, if a user selects a specific location from the pinned ones, a pop-up appears and gives a summary of the analysis results, as shown for Pilot Larisa in Figure 25.

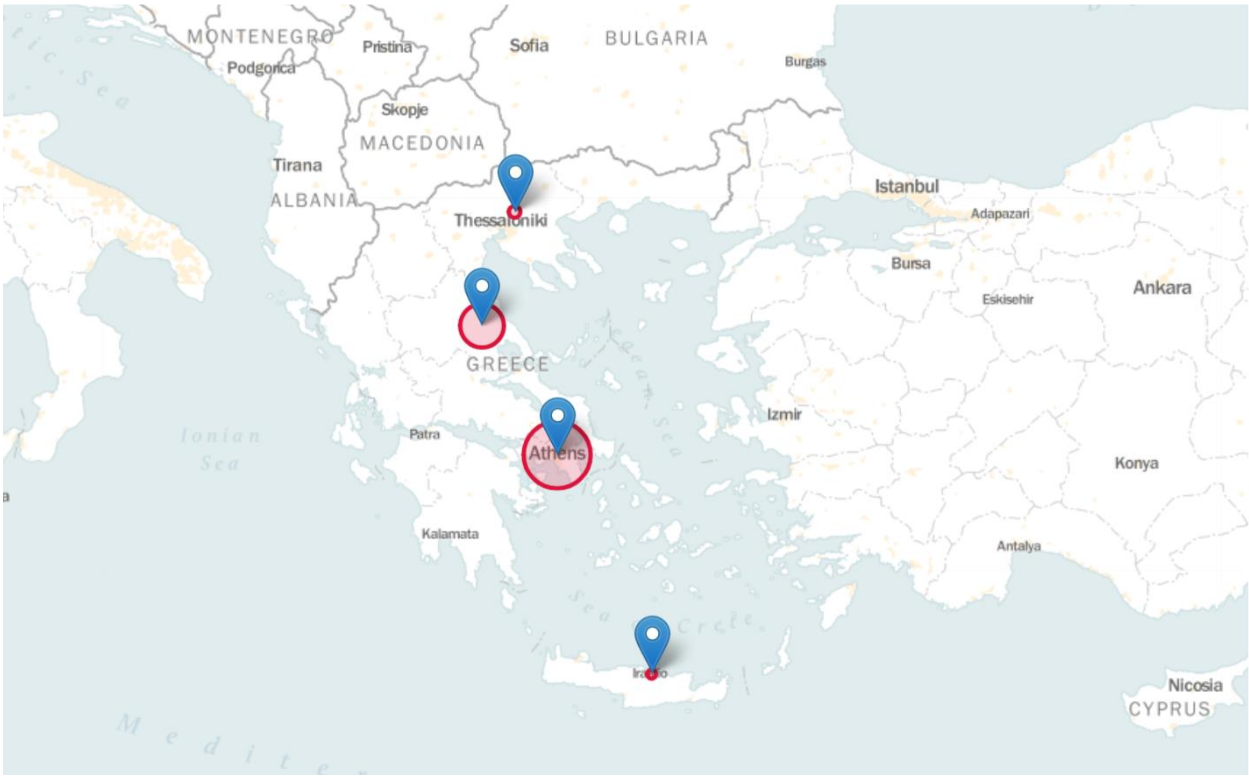
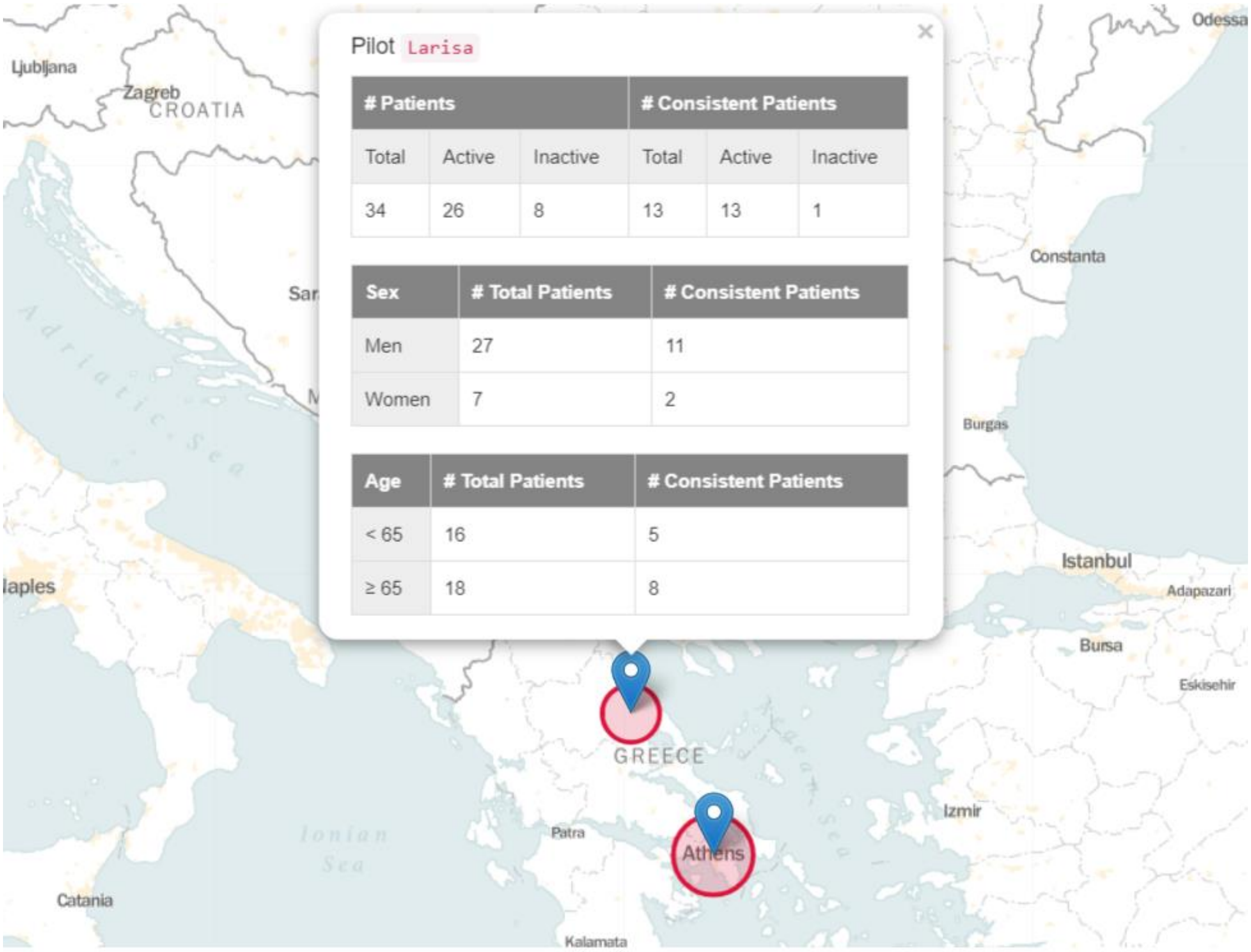


Figure 24: “Map of Pilots in Greece” User Interface



Pilot Larisa

# Patients			# Consistent Patients		
Total	Active	Inactive	Total	Active	Inactive
34	26	8	13	13	1

Sex	# Total Patients	# Consistent Patients
Men	27	11
Women	7	2

Age	# Total Patients	# Consistent Patients
< 65	16	5
≥ 65	18	8

Figure 25: Pilot Larisa – Map User Interface

5.3.4 Distinguish “Good” – “Bad” Days

At this point of the analysis, we have resulted in twenty-one patients that follow a regular schedule in performing the daily measurements of their biomedical signals. With the aim of exploring the dataset further and extracting as much information as we can, we decided to study, not the frequency of the measurement – taking this time, but the measurements themselves. In other words, we are aware of the fact that the dataset is comprised of patients, but how are the values of their measurements distributed on a daily basis? Is there a fluctuation of these values? Is there an improvement in the medical condition of the patients and which factors affect such a variance, if any?

5.3.4.1 Anomaly Detection

In order to answer the above questions, we started the analysis procedure by discovering and visualizing the outlier values of each one of the measurements. With the use of z-score and the boxplot visualization method, we were able to detect possible anomalies of the measurement values and, in combination with Table 11, decide which values are proven to be abnormal in living organisms and have most likely been provoked by devices’ malfunctions or external factors. As an alternative, we also performed the Chauvenet’s Criterion to check if there are any outlier values, but the results were similar to the ones of z-score.

The following pictures present the statistical results that we were given:

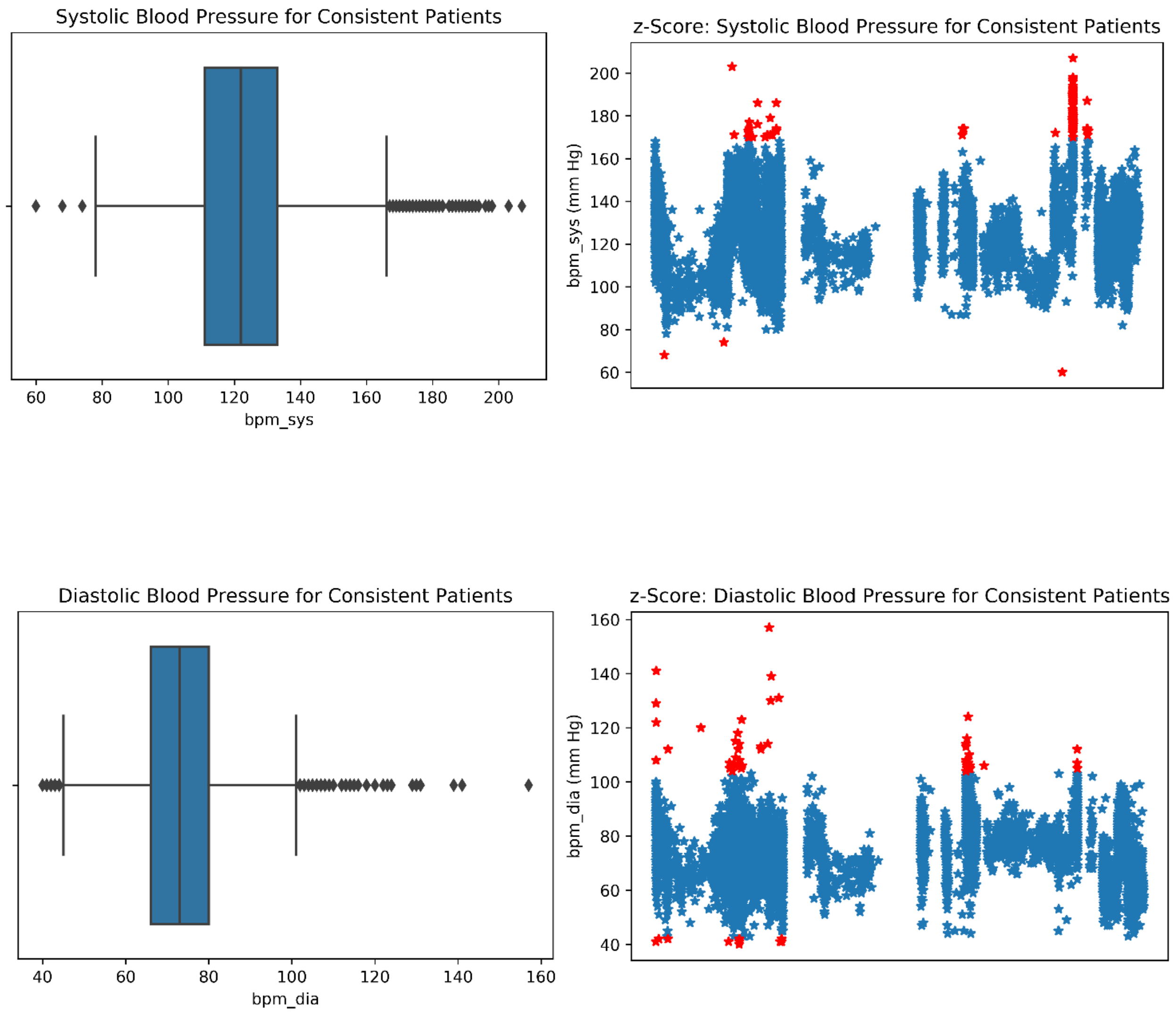


Figure 26: Anomaly Detection – Blood Pressure

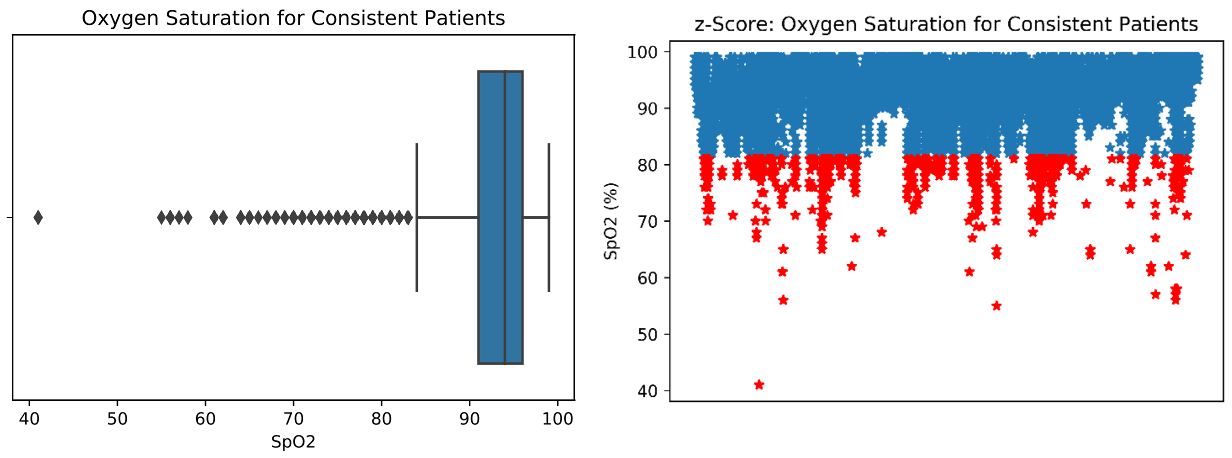


Figure 27: Anomaly Detection – Oxygen Saturation

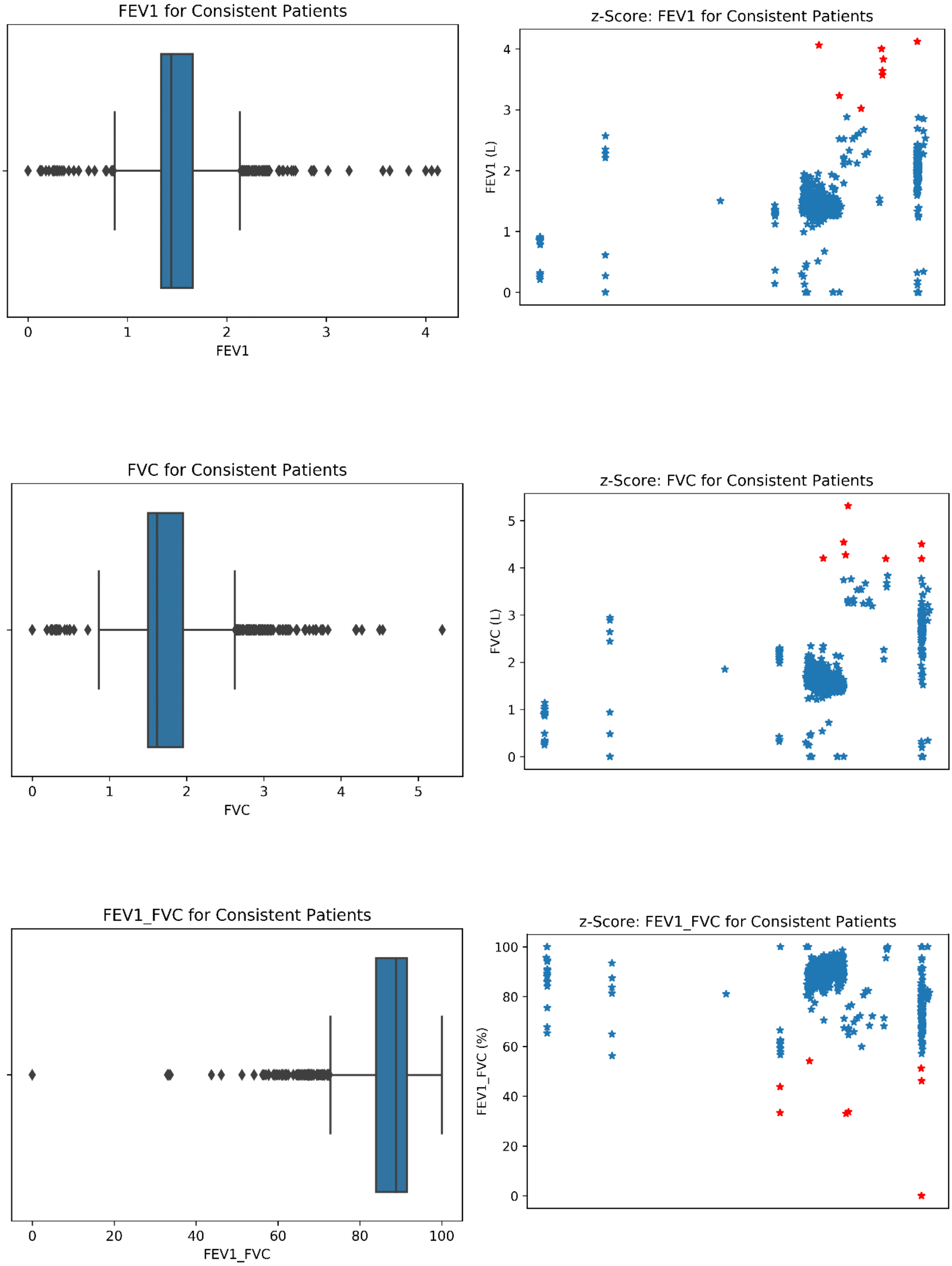


Figure 28: Anomaly Detection – FEV1, FVC, FEV1/FVC

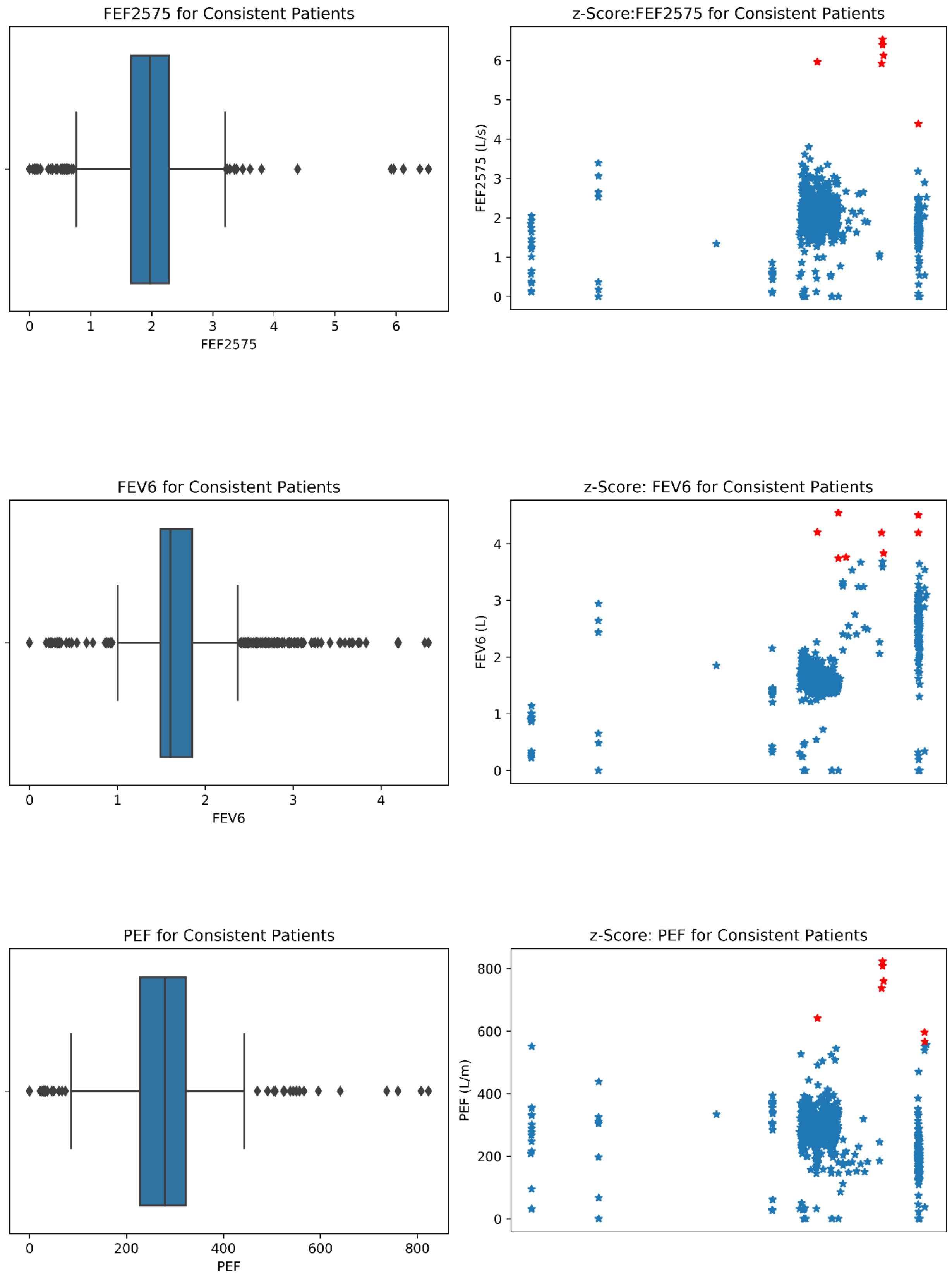


Figure 29: Anomaly Detection – FEF2575, FEV6, PEF

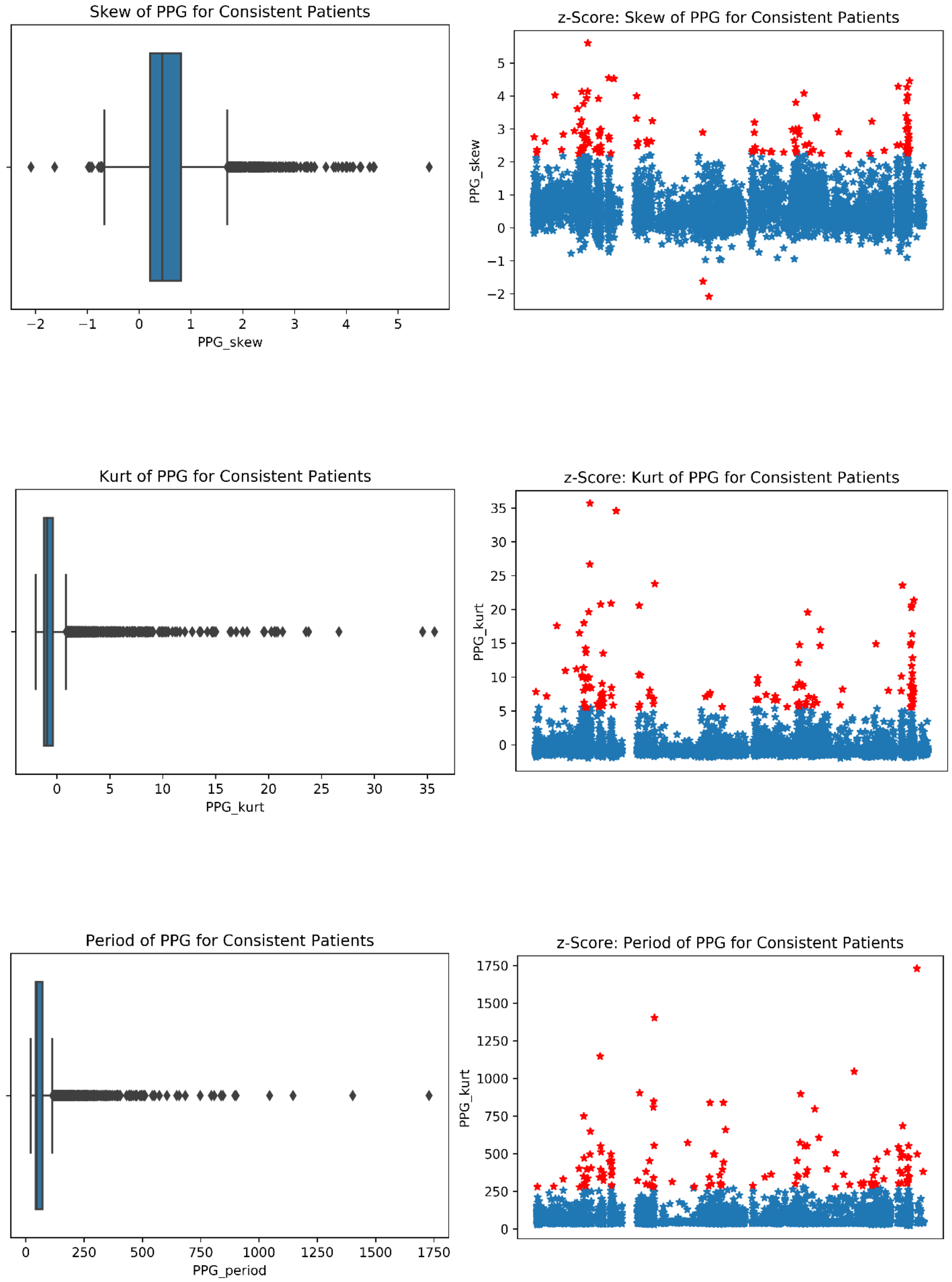


Figure 30: PPG_skew, PPG_kurt, PPG_period

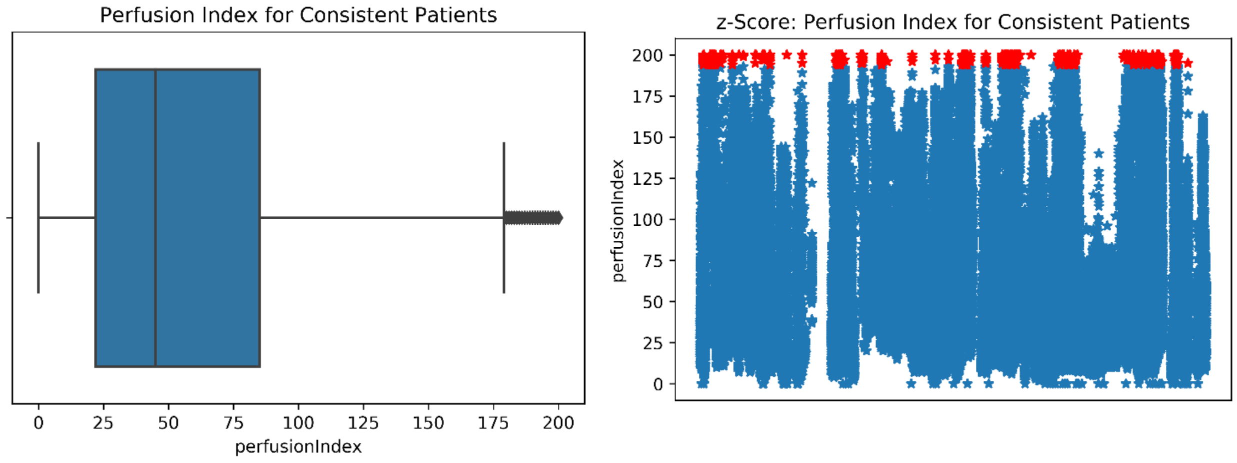


Figure 31: Anomaly Detection – Perfusion Index

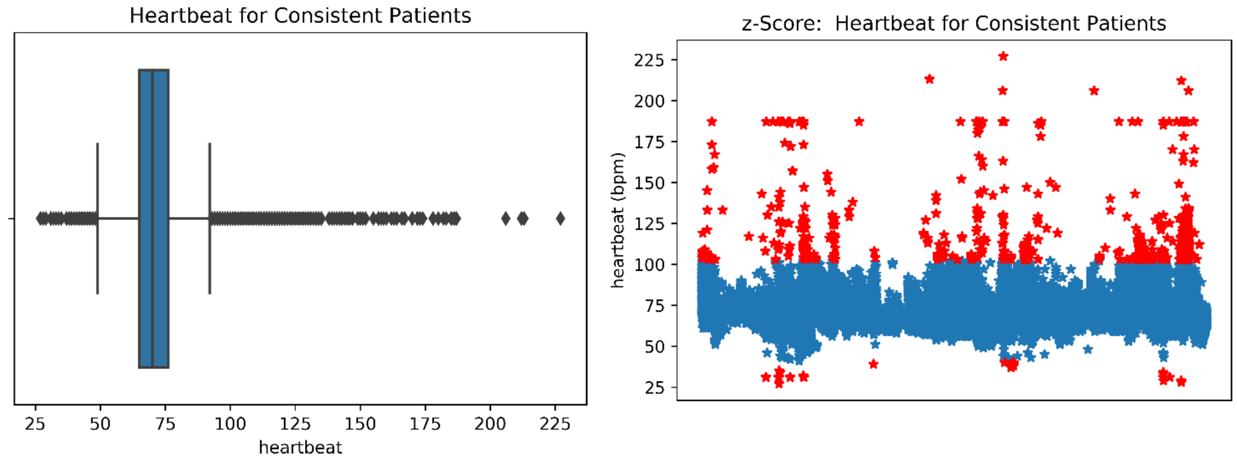


Figure 32: Anomaly Detection – Heart Rate

Taking into consideration the interdisciplinary nature of healthcare analytics, domain expert opinion was essential and constituted integral part of the anomaly detection procedure. More specifically, the following table summarizes the minimum and maximum medically allowed values that a patient, despite the severity of their medical situation, could have under normal measurement conditions:

Table 11: Normal Biosignal Values

BioSignals	Values		Ideal Medically Accepted Values for Chronic Patients	Maximum Medical Limits for Extreme Situations	z-Score min – max Values
Blood Pressure	**Systolic**	**Diastolic**			
Low (Hypotensive)	< 80	< 60	Systolic 90 – 120	Systolic 70 – 190	Systolic 76 – 170

Normal	**80 – 120**	**60 – 80**			
Pre-Hypertensive	120 – 139	80 – 89	Diastolic 60 – 80	Diastolic 40 – 100	Diastolic 42 – 104
High (Hypertensive)	> 140	> 90			
Oxygen Saturation					
Severely Hypoxic	< 85 %		88 – 100 %	85 – 100 %	83 – 100 %
Hypoxic	85 – 94 %				
Normal	**≥ 95 %**				
Heartrate					
Excellent	< 61		60 – 80	40 – 100	41 - 103
Good	**62 – 67**				
Average	68 – 80				
Poor	> 75				
Spirometry					
FEV1	80 – 120 %		58 – 67 %	51 – 60 %	FEV1/ FVC 56 – 116 %
FVC	80 – 120 %		59 – 65 %	52 – 58 %	
FEV1/FVC	95 – 105 %		83 – 87 %	75 – 80 %	
FEF25-75%	> 79 %		56 – 66 %	49 – 59 %	

In the first column of values, in Table 11, we can see the range in which the normal values of a persons' measurements should be, according to medical bibliography. In the second one, we present the range of values that are usually measured in cases of chronic patients and mostly of those suffering from IPF. The third column of values presents the minimum and maximum values that medical experts have noticed in very extreme and severe cases, according to their testimony. Last but not least, in the last column, we added the values that were given from the statistical z-Score test.

Comparing the medical evidence with the statistical results, we can see that the value limits are almost identical in very severe cases. Consequently, we decided to remove from the dataset the values with $z - Score > 3$ and $z - Score < 3$ since they are out of the allowed medical range, in order to continue with the analysis. The following table presents the total number of abnormal, outlier values that were removed from the dataset:

Table 12: Total No. of Abnormal Values

Bpm_sys	107	FEV1/FVC	8
Bpm_dia	61	FEF25-75%	6
Heartbeat	2166	FEF2575	6
SpO2	2851	PEF	7
Perfusion Index	1829	PPG_kurt	123
FEV1	8	PPG_skew	125
FVC	7	PPG_period	127

5.3.4.2 Data Overview

Before proceeding to the next steps, we needed to have an insight of the data that we kept, thus, we plotted the number of total reading values that all the patients take, per day, and the number of total reading values per reading category:

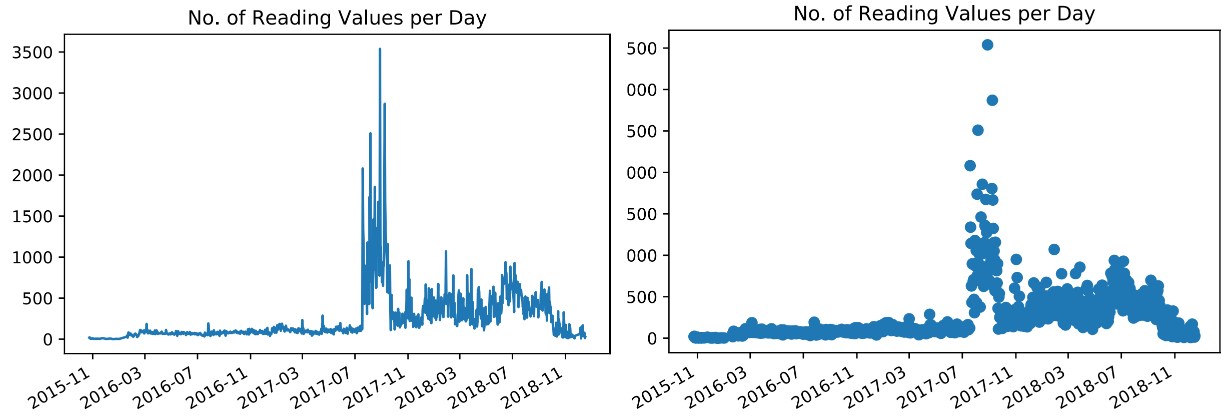


Figure 33: No. of Reading Values per Day After Removing the Ubnormal Values

Total number of reading values per reading category, performed by consistent patients.

FEF2575	267320	SpO2	90251
FEV1	267322	heartbeat	79467
FEV1_FVC	267332	perfusionIndex	111487
FEV6	267321	PPG_united	260360
FVC	267321	PPG_skew	260885
PEF	267321	PPG_kurt	260883
bpm_sys	257069	PPG_period	260887
bpm_dia	257023		

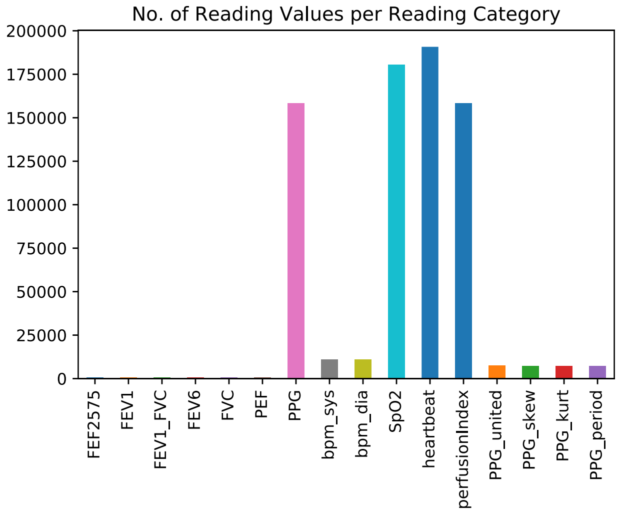


Figure 34: No. of Reading Values per Reding Category After Anomaly Removal

5.3.4.3 Correlations

Observing the above plots, we can see that it would be crucial to study how these different types of measurements interact with each other. Therefore, we calculated the mean value for every biosignal, per day, for each patient – as presented in Figure 35 – and we made some plots that illustrate the correlations among the different measurement categories. What attracts our interest is Figure 36 that is a heatmap visualization of the correlations between the reading categories per day, where the more close to 1 the represented values are the more correlated the reading categories, and, Figure 37 that presents these correlations in a scatter plot, starting from the 1st day when a PPG measurement took place.

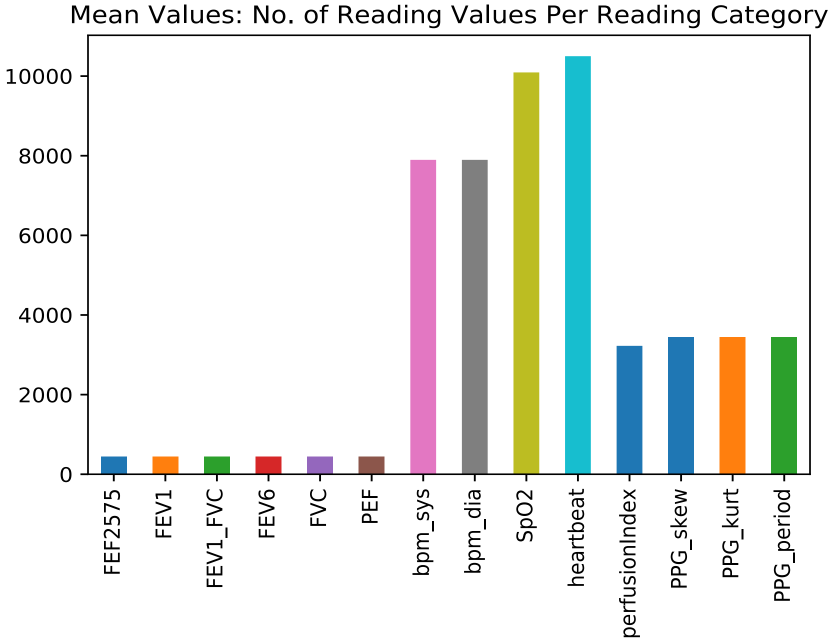


Figure 35: No. of Reading (Mean) Values per Reading Category

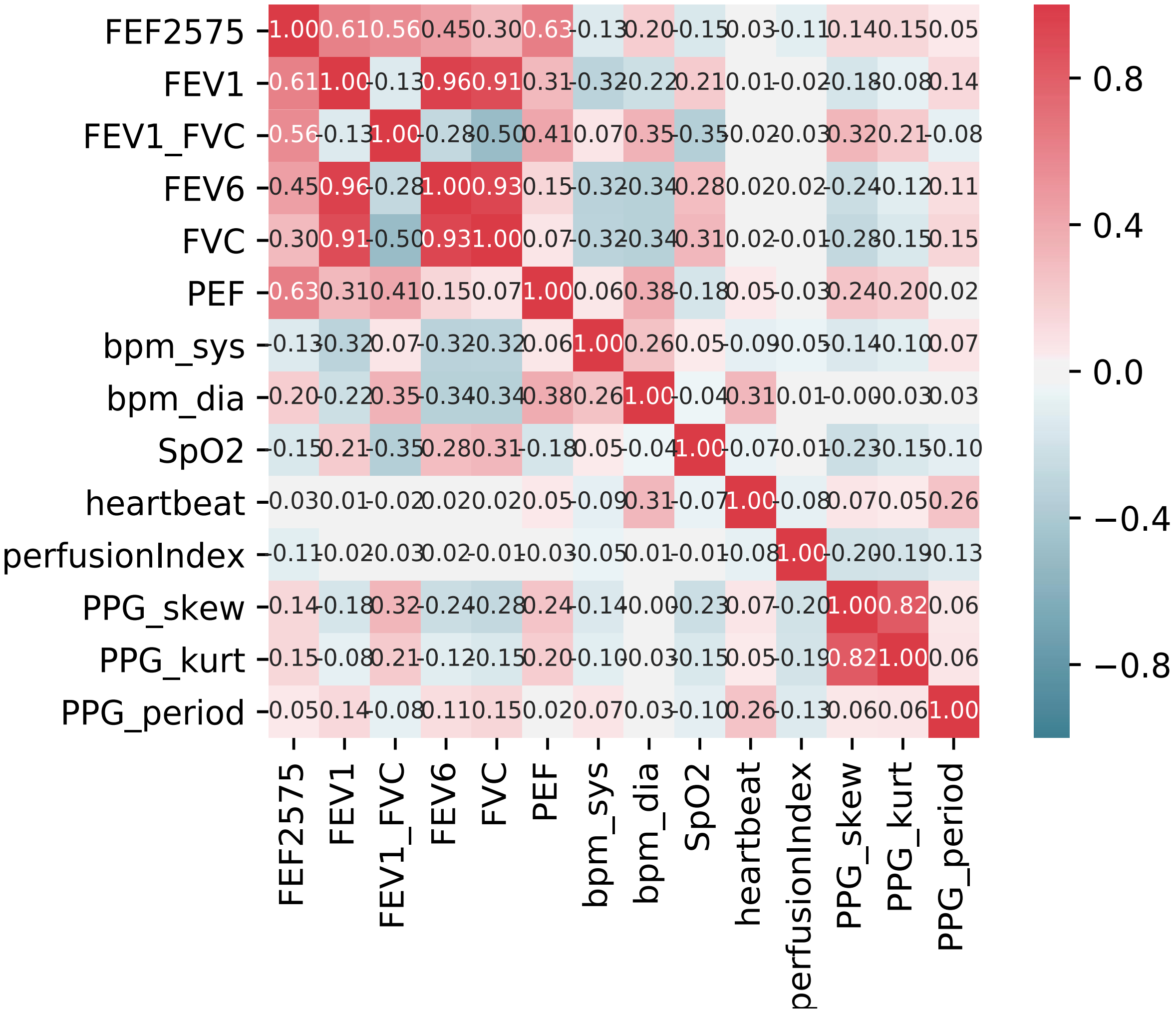


Figure 36: Heatmap – Correlations Among Biosignals

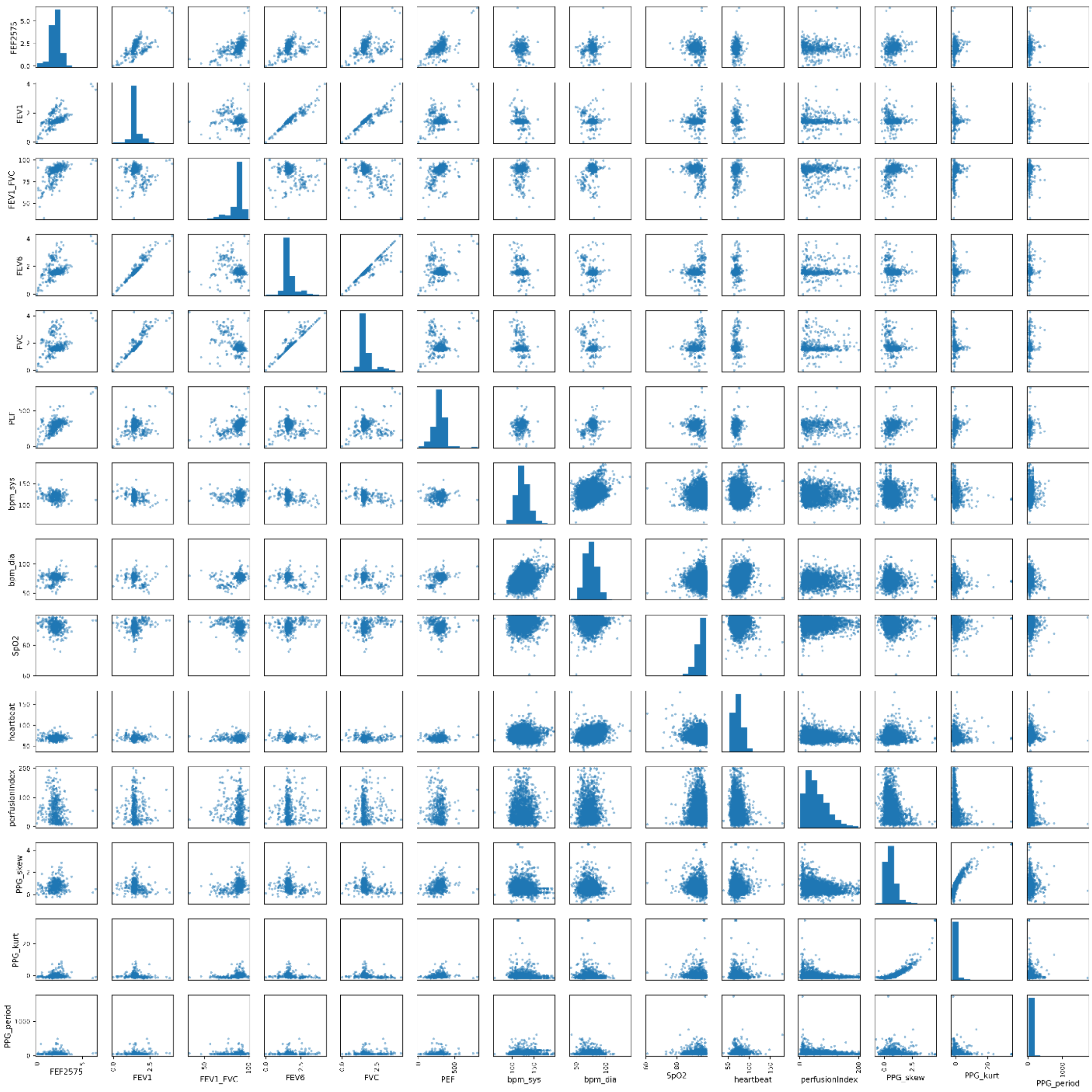


Figure 37: Correlations Among Biosignals

From the above two figures, we can observe that the PPG derivatives, PPG_skew and PPG_kurt, are highly correlated as well as all the resulting values from spirometry measurements, which seem to interact with each other intensively. The most noticeable of the rest of the correlations can be summarized in the following table:

Table 13: Correlated Reading Values

Medium Correlation	**Low Correlation**	
bpm_dia – heartbeat	bpm_sys – SpO2	Heartbeat – PPG_period
bpm_dia – PEF, FEV1/FVC	bpm_sys – PEF, FEV1/FVC	SpO2 – FEV6, FEV1
SpO2 – FVC	bpm_dia – FEF2575	PPG_skew – FEF2676, PEF
PPG_skew – FEV1/FVC	PPG_kurt – FEF2575, PEF, FEV1/FVC	PPG_period – FVC, FEV1, FEV6, FEF2575
bpm_sys – bpm_dia		

Knowing almost everything about the data and the relations among them, we needed to proceed with the analysis. First and foremost, we decided to exclude from the dataset the two participants from Pilot Athens, since, we don't really know if they are patients or not and we don't want the clustering results that will follow, be altered.

5.3.4.4 Handling Missing Values

In order to prevent any other complications during the clustering procedure, we needed to check if there were any missing values that needed to be fixed. For this purpose, after performing several techniques, we decided to impute the mean value from the existent 15 nearest neighbor measurement values, implementing the k-Nearest Neighbor algorithm.

Hereafter, having a complete dataset, we were able to continue with the clustering procedure.

5.3.4.5 Clustering

Our goal was to distinguish the days when a patient's health had a positive lead from the ones when the patient's medical condition was deteriorating. In other words, we needed to cluster the patient's everyday life into two categories, the "Good" days and the "Bad" ones.

5.3.4.5.1 Standardization

As it has already been explained, it is necessary to standardize the data before implementing the algorithm, due to its sensitivity. Therefore, based on the same technique and the theory of the paragraph 3.3.6, we calculated the mean value and the standard deviation, and we standardized our data.

5.3.4.5.2 Clustering with k-Means

Moving on to the algorithm, we set up a two-cluster k-Means problem and, when we applied it to the data, we got the results that are summarized in Table 14.Observing the results, we noticed that the centers of the two clusters do not differ significantly, a rather surprising fact, since our sample consists of patients suffering from a chronic condition. However, we were able to identify slight superiority of the values in Cluster 1, thus, we understood it represents the "Good" days:

Table 14: "Good – Bad Days" Cluster Centers

Reading Category	Cluster 0	Cluster 1
bpm_sys	125.455046	122.624274
bpm_dia	75.451670	72.664638
SpO2	94.616105	96.617773
heartbeat	72.146020	74.876081
perfusionIndex	56.993308	54.569338
FEF2575	1.988098	2.102237
FEV1	1.470863	1.881505
FEV1_FVC	86.465055	81.658685
FEV6	1.656691	2.202137
FVC	1.722978	2.362825

PEF	290.100331	272.153291
PPG_skew	0.542069	0.472126
PPG_kurt	-0.612700	-0.747887
PPG_period	63.332784	65.828770

5.3.4.6 Dimensionality Reduction for Visualization

In order to inspect the results visually, we exploited the capabilities of t-SNE that reduces the features' dimensions to only two, and we got the following visualization, where we can see how well k-Means performed in separating the data into the two clusters:

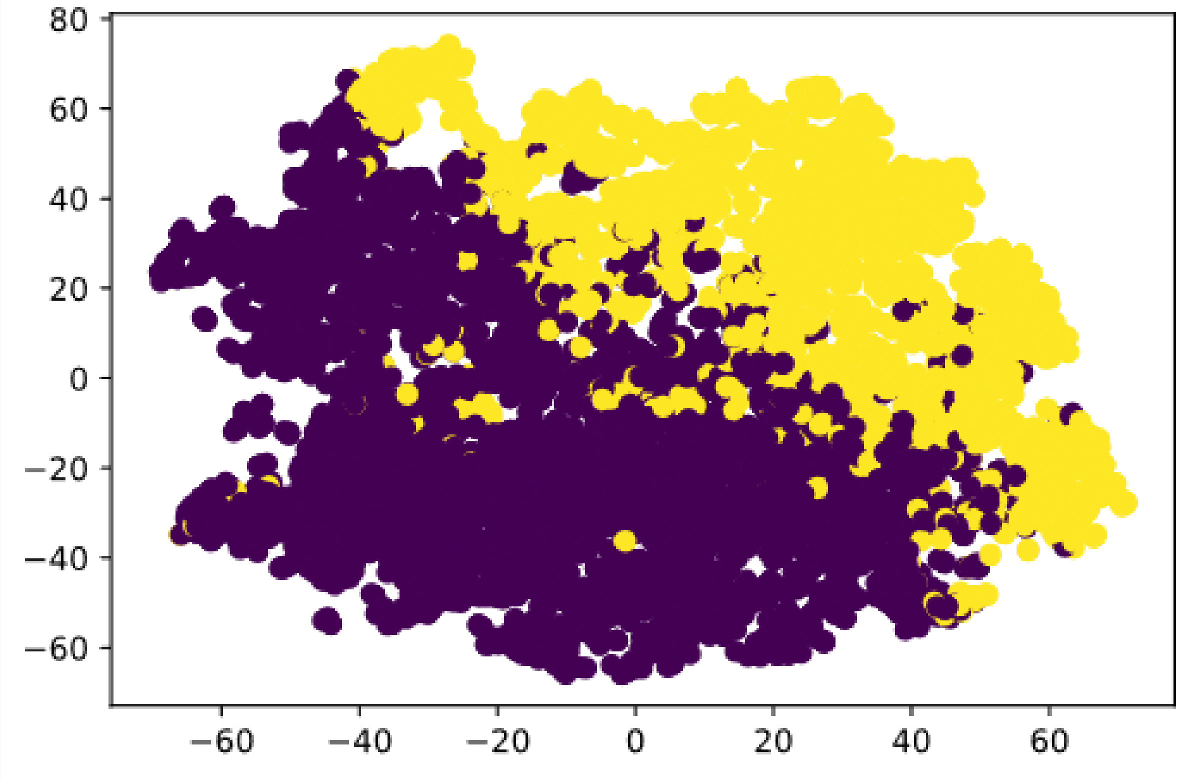


Figure 38: Good and Bad Days Results – Visualization with t-SNE

5.3.4.7 Results Communication

Within the framework that all these data have been generated from the usage of a mobile application, we wanted to create a visualization of the clustering results that could be readable and potentially useful to the end user.

For this purpose we extracted, indicatively, the clustering results for a particular patient during a random month, July 2018. Using the "Sunburst" visualization tool of MS Excel, and giving to it the clustering results as input, we were able to get the following figure that could be considered as a user-friendly interface of the mobile application:

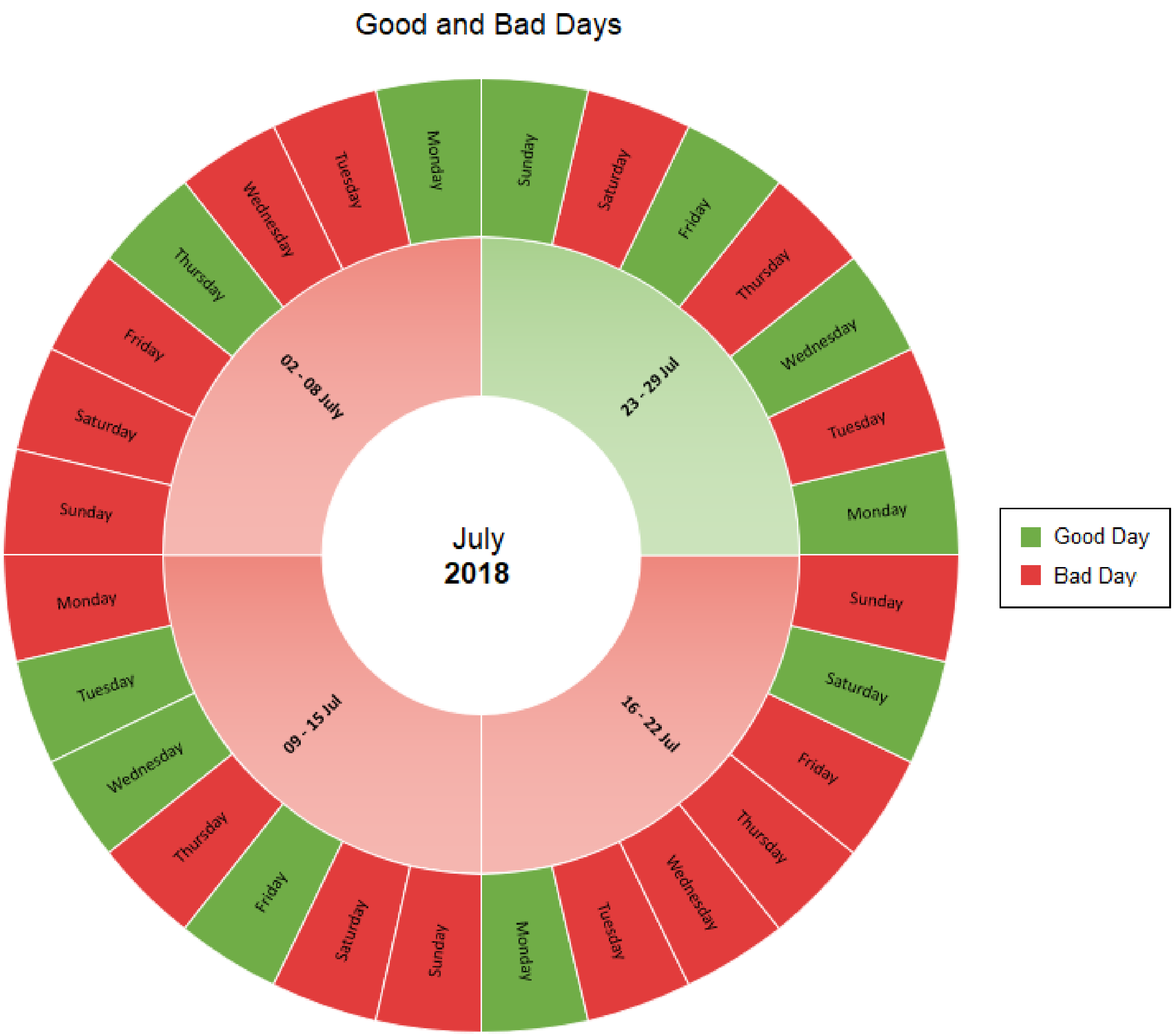


Figure 39: "Good – Bad Days" User Interface

5.3.5 Statistical Inference

At this point, the analysis procedure has been completed successfully and the data have undergone multiple processing techniques, a fact that makes them appropriate candidates for a statistical t-test. Since we are given the answers to the questions that we posed in the beginning, it would be very interesting to compare the results with the demographic information of the dataset, the age and the sex, in order to understand per population the statistical significant differences in the medical values.

For this purpose, we selected the Welch’s t-test as it can function even with data of different variances. Since its prerequisite is that the data follow the normal distribution, we took the dataset that occurred after anomaly detection and outlier values handling in the previous section and performed the Kolmogorov-Smirnov test. All the parameters (columns of the dataset) were found to be following the normal distribution, so, we continued and performed Welch’s t-test in order to check if these columns differ significantly among the populations of different geographical areas, different age groups and between the two sexes.

The results can be presented in the following tables and figures that show both the statistical differences among the variables and the normal distribution they follow:

Table 15: Welch' s t-test results: Men – Women

Hypothesis: The mean per measurement is the same between men and women		
Measurement Category	**P-value**	**Statistical Difference**
FEF2575	0.0001053367445997203	Yes
FEV1	1.1490484084858517e-10	Yes
FEV1_FVC	0.1333608986088117	No
FEV6	0.004480605511758619	Yes
FVC	0.02257680016241423	Yes
PEF	0.5645004423431178	No
bpm_sys	9.734630324196489e-84	Yes
bpm_dia	9.380079191815203e-12	Yes
SpO2	3.8436912833536457e-287	Yes
heartbeat	0.0	Yes
perfusionIndex	0.0	Yes
PPG_skew	4.231615010004724e-21	Yes
PPG_kurt	1.70276504713253e-10	Yes
PPG_period	8.367322403682799e-36	Yes

Indicatively we present the following graphical results:

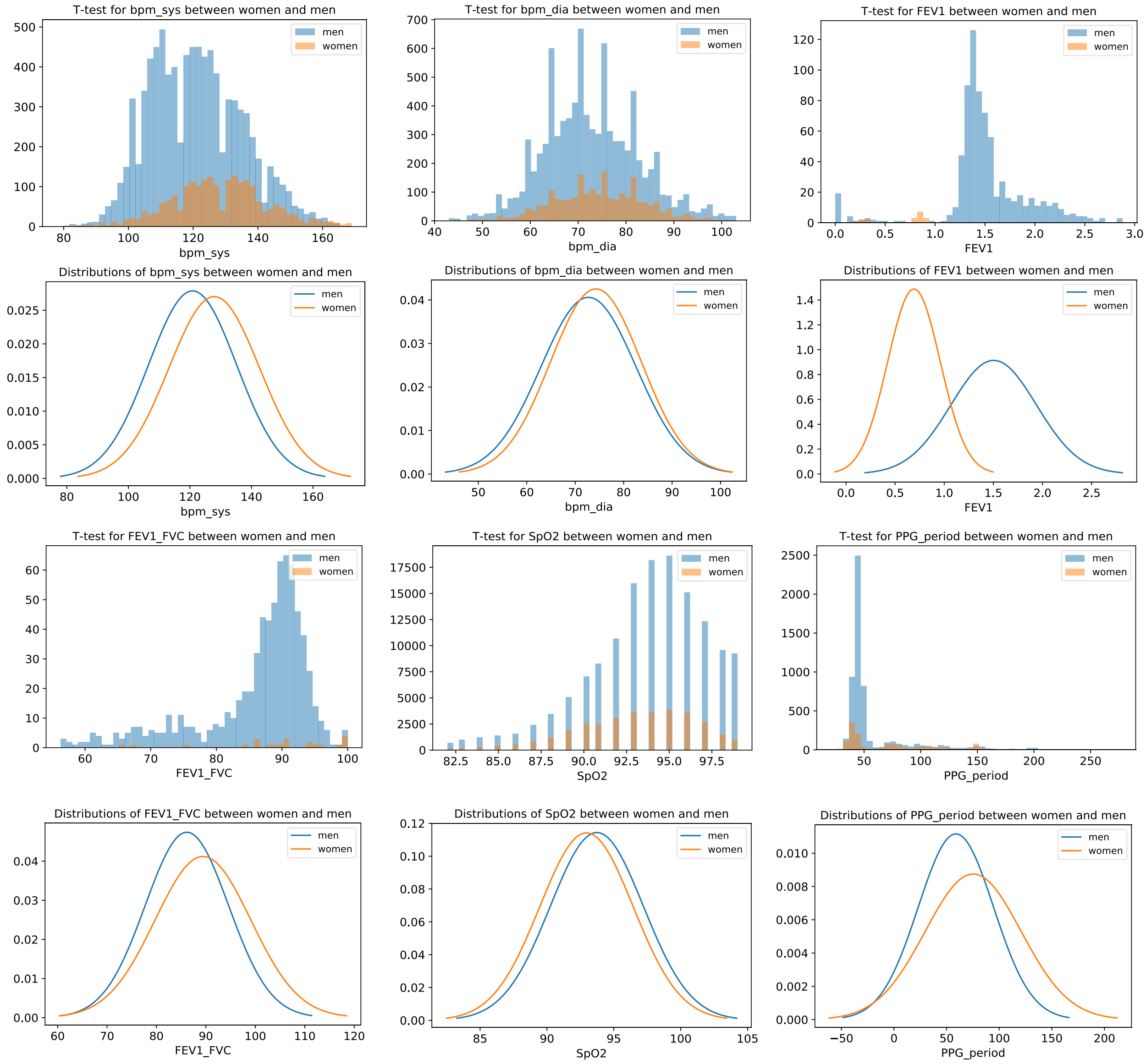


Figure 40: t-test and Distributions for Men and Women

Table 16: Welch’ s t-test results: Age <65 – Age >= 65

Hypothesis: The mean per measurement is the same for patients aged over 65 years old and under 65 years old		
Measurement Category	**P-value**	**Statistical Difference**
FEF2575	3.484270067112345e-16	Yes
FEV1	2.7718653572184592e-06	Yes
FEV1_FVC	1.0407113206836113e-35	Yes
FEV6	1.0142242351454229e-13	Yes
FVC	1.1797603332039976e-18	Yes
PEF	1.0062283696451662e-20	Yes
bpm_sys	1.9084819597458553e-38	Yes
bpm_dia	0.0	Yes
SpO2	0.0	Yes
heartbeat	1.0174592950556485e-35	Yes
perfusionIndex	0.0	Yes
PPG_skew	7.676253501010304e-50	Yes
PPG_kurt	8.840204601984984e-18	Yes
PPG_period	0.09950856455428282	No

Indicatively we present the following graphical results:

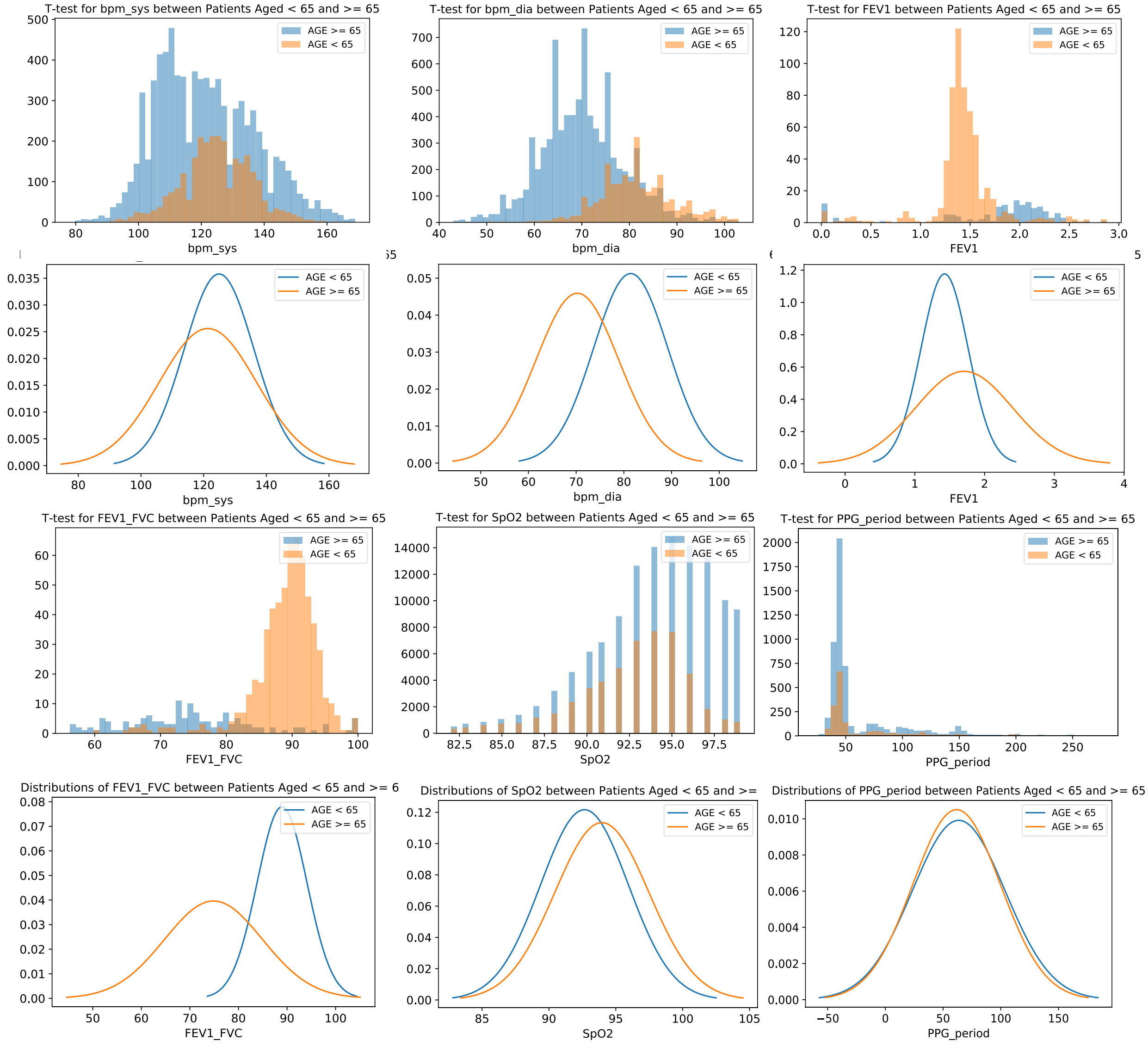


Figure 41: t-test and Distributions for Patients Aged <65 and Aged >= 65

Table 17: Welch' s t-test results: Pilot Athens – Remote Pilots

Hypothesis: The mean value of the measurements of patients living in an urban city (Athens) is the same with those who live in remote areas (remote pilots)		
Measurement Category	**P-value**	**Statistical Difference**
FEF2575	Could not be calculated	-
FEV1	Could not be calculated	-
FEV1_FVC	0.000585137334004672З	Yes
FEV6	0.0028835164245З0072	Yes
FVC	0.0004197722727246886	Yes
PEF	Could not be calculated	-
bpm_sys	5.52040558141934e-12	Yes
bpm_dia	1.7092178868002557e-06	Yes
SpO2	3.554723340192662e-56	Yes
heartbeat	0.0	Yes
perfusionIndex	0.0	Yes
PPG_skew	3.4519548397546584e-46	Yes
PPG_kurt	1.568743384818745e-21	Yes
PPG_period	4.521300021168656e-29	Yes

Indicatively we present the following graphical results:

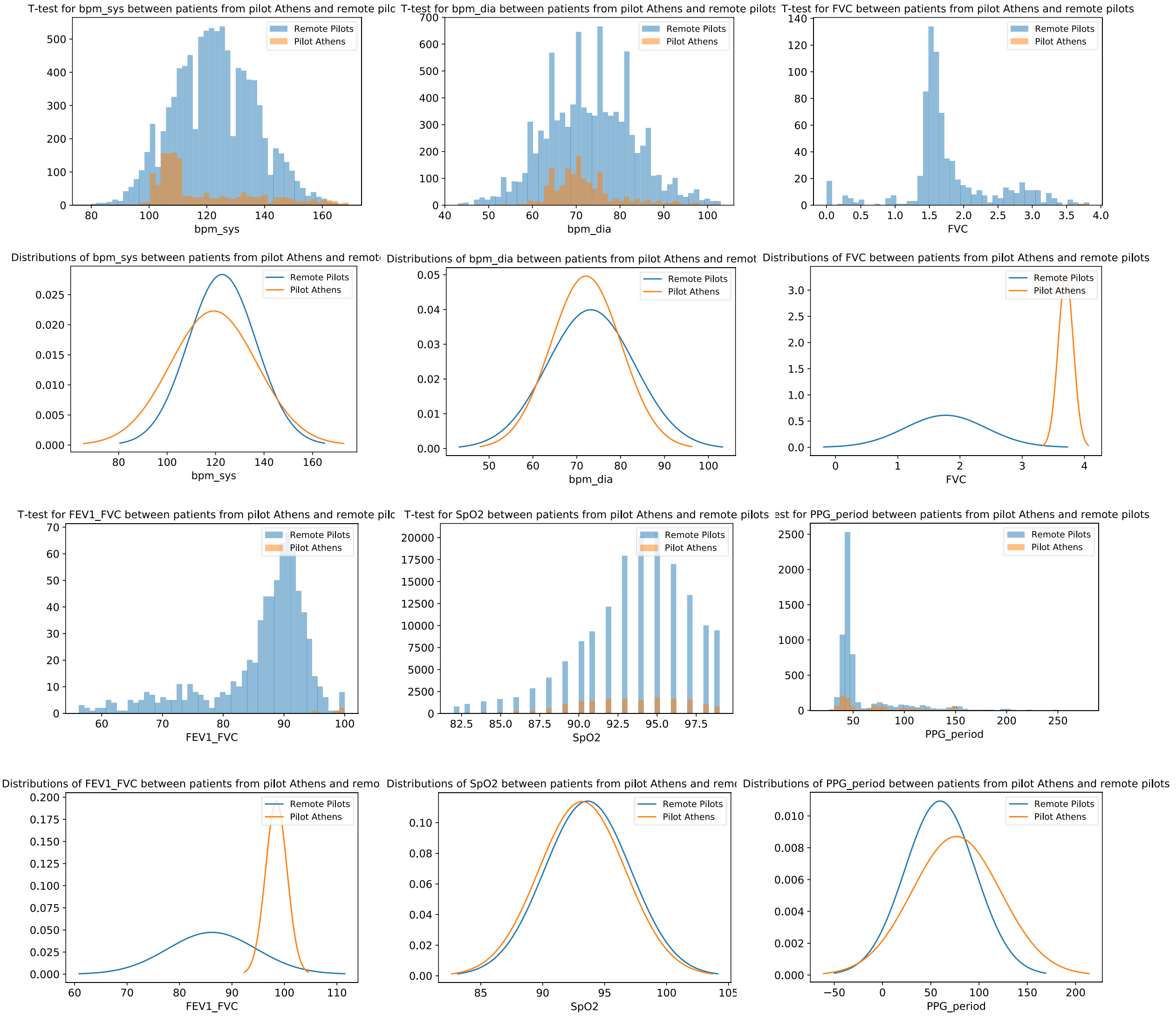


Figure 42: t-test and Distributions for Pilot Athens and Remote Pilot Patients

6. IMPLEMENTATION

6.1 Languages

The programing language that has been used for the processing of the data and their analysis, in order to fulfil purposes of the current thesis project, is Python. Python was "born" in 1990 [61] as a language with main aim its ease of use and is distinguished by its many libraries that make it quick and easy to learn. Nowadays, Python constitutes one of the most common, high-level, object-oriented programming languages in data science. Its design philosophy emphasizes in readability, as its syntax allows developers to write much less code than ever.

HTML has also been used in order to fulfill the needs of the web – based map visualization. Developed in 1990 as well, this HyperText Markup Language (HTML) is a computer standard markup language used for creating web pages and web – based applications. It is very simple in use and, along with CSS and JavaScript, "forms the triad of cornerstone technologies for the World Wide Web" [62].

6.2 Libraries

Python's available libraries guarantee the language's success. Some of the most important ones, used in the present implementation, are:

6.2.1 NumPy

NumPy is one of the first libraries behind Python's success story and constitutes the fundamental package on which all the essential higher-level tools for scientific computing are built. Among the features it provides, the most important are: the N - dimensional array, a multidimensional matrix with fast and efficient memory that provides numerical operations, the standard mathematical operations, such as linear algebra or Fourier Transformations [63], among others, and the fact that it provides a very easy way to transfer data to external libraries written in lower level languages or to external libraries that return data to Python as NumPy arrays.

6.2.2 Pandas

Pandas is the Python data analysis library, used for everything, from importing data from Excel spreadsheets, to processing datasets for time series analysis. It provides almost all the common tools that can be used for data handling, which means that powerful Pandas data frames can perform data pre-processing and clearing. Pandas is built on NumPy and it contains high-level data structures and tools designed to make data analysis fast and easy [64]. Using Pandas, it's easier to handle the missing data, it merges other relational functions found in popular databases and it constitutes the best tool for data wrangling.

6.2.3 SciPy

The SciPy library is also depended on NumPy, which provides easy and fast N - dimensional array manipulation. It is designed to work with numerical NumPy tables and provides many user-friendly and efficient numerical routines, such as routines for numerical integration, modules for optimization, linear algebra and other common tasks of data science [65].

- from scipy.**cluster** import **hierarchy** [66]

We used the cluster.hierarchy to perform Hierarchical Clustering with the **'complete' linkage** method. The clustering results were plotted with **dendrogram** and in order to form flat clusters from it and interpret the results, we used the **fcluster** with the **maxclust** parameter in order to have 2 clusters.

- from scipy.**signal** import **fftconvolve**

Used to convolve the smoothed PPG signal with the Fast Fourier Transform method, in order to find the autocorrelation.

- from scipy.**signal** import **savgol_filter**

We applied the Savitzky-Golay filter to smooth the PPG signals and, as a result, repair the false small peaks.

- from scipy.**stats** import **kstest**

We performed a Kolmogorov-Smirnov test to check if each one of the dataset's features' distribution follows the normal one.

- from scipy.**stats** import **ttest_ind**

We performed a two-sided t-test for the null hypothesis that two independent samples have identical mean values. For this purpose we used the Welch's t-test which does not assume equal population variance. The null hypothesis was rejected for a $p < 0.05$.

- from scipy.**stats** import **norm**

We produced the normal distributions of the different populations to visualize the results of the t-test.

6.2.4 Scikit – learn

Scikit-learn constitutes a library dedicated for machine learning purposes and it is based on NumPy and SciPy. This library contributes to the rapid implementation of known algorithms in the dataset and includes tools for many typical machine learning and data mining tasks, such as grouping, sorting, regression, dimensionality reduction, clustering etc. [54].

- from sklearn.**cluster** import **KMeans**

This module, along with the fit () function, enabled us to perform clustering of unlabeled data using the k-Means algorithm.

- from sklearn.**manifold** import **TSNE**

We used the fit_transform () method in order to fit the tested dataframe into an embedded space and get it back transformed.

6.2.5 Fancyimpute

Fancyimpute is a data mining library that provides advanced capabilities in imputing the missing values.

- from fancyimpute import **KNN**

We used this module to take advantage of the nearest neighbor imputation capability, which weights samples using the mean squared difference on features, for which two rows have observed data [67]. In our case, we used $k = 15$ nearest neighbors which we believe that is not too small, so that noise would have a higher influence on the results, nor too large to make it computationally expensive.

6.2.6 Math

The python module that provides access to simple mathematical functions.

We used this module to calculate the standard deviation of the two samples during statistical inference experiments.

6.2.7 Matplotlib

Matplotlib is a plotting library used for visualization purposes. It allows different types of plots, such as line graphs, pie charts, histograms, etc. and has the ability to support interactive features and environments, different GUIs and can export the visualizations to common vector and graphic formats, like PDF, SVG, JPG, PNG, etc. [68].

- from matplotlib import **pyplot**

Pyplot was used for line charts, scatter plots and histogram visualizations and is responsible for fitting the data to the relevant graphical representation.

6.2.8 Seaborn

Seaborn is a visualization library based on Matplotlib and it provides attractive and informative drawings of statistical graphs, in a high-level interface, like boxplots, heatmaps and so on.

We used seaborn in order to create **boxplot** visualizations.

6.2.9 Folium

Folium constitutes another data visualization library which mainly focuses in geospatial data representations. The data can bind to an interactive map and be used for choropleth visualizations.

With folium, we were able to create a folium.**Map** of the pilot programs based on the latitude and longitude values of the respective cities that they took place

6.3 Jupyter Notebook

The Jupyter notebook constitutes an interactive computing approach, providing a qualitative web-based application to fulfil all the needs of a computation process, including development, documentation, code execution and results communication.

The analysis procedure that we performed has been executed in the environment of the Jupyter Notebook.

6.4 GitHub Environment

GitHub Inc. is a web-based service, mostly used for hosting computer code and It offers the git functionalities of version control and source code management. The source code of the project that we have implemented is available in GitHub under the repository: https://github.com/katerinag19/master-thesis.

6.5 Pseudocode

In this section we provide in pseudocode the procedure we followed in order to analyze the PPG signal and extract some valuable and measurable features from it, and the one we followed to extract the appropriate features to be used as input for the clustering algorithms.

6.5.1 PPG Signal Analysis

Append consecutive PPG signal values and extract features:

```
For each patient:
    For every day:
        For every i PPG record:
            If the time difference between the PPG record i and the PPG record i+1 is maximum 2 seconds:
                Append PPG records
            Else:
                Smoothen the – so far – united PPG
                Find the autocorrelation of the united PPG
                Calculate the distance between the first two peaks in the autocorrelation to find the period
                Calculate Skewness of the united PPG
                Calculate Kurtosis of the united PPG
```

6.5.2 Feature Extraction for Clustering

Extract the features: 1. mean of day differences, 2. total days, 3. total readings, 4. total days per total readings:

```
For each patient:
    If they have at least one record per day:
        Find how many days mediate between two consecutive measurements
        Get the mean of these days (1st feature)
        Get the sum of days from the 1st measurement to the last (2nd feature)
        Get the sum of the total measurements (3rd feature)
Produce the ratio total days/ total measurements (4th feature)
```

7. CONCLUSIONS AND FUTURE PERSPECTIVES

The constant monitoring and management of chronic patients' health condition is undeniably a vital need and one of the greatest challenges of modern healthcare systems. The evolution in the field of mobile health with the integration of various sensor technology devices has made the Internet of Things an integral part of the patients' daily life, as it monitors almost every aspect of it and gathers enormous quantities of data. In the previous sections, we presented the scientific background that has led to the development of such innovative electronic healthcare services dedicated to elder patients and, based on a – data science – literature review, the analytics techniques and methodologies that could be applied on such "Big Data", to handle them and extract the most valuable information out of them. The main purpose of the current thesis project was to take advantage of the theoretical background and, based on the case study of BioAssist, to propose a data analysis solution that could control these data, understand them and drive conclusions that could be used as beneficial feedback both to stakeholders and the patients, in order to enhance the prosperity and longevity of the latter.

7.1 Conclusions

Making a brief recap of the procedure we followed, we started off with the preprocessing of the data, performing various techniques. Noticing that one of the main differences among the patients is the frequency of measuring their biosignals, we tried to distinguish them into consistent and inconsistent, following the procedure of clustering. Proceeding with the consistent patients, we examined the quality of their measurements and we identified values' abnormalities. Keeping the normal values, we studied them and tried to characterize the overall daily performance of each patient as good or bad, depending on the correlations among the different measurement categories. Last but not least, we compared the results with the demographic information of the dataset, the age, the sex and the patients' geolocation, in order to investigate the statistical significant differences in the medical values, per population.

The results indicated that the more consistent a patient is, the better chances of chronic disease survival they have. Mathematically, only 1 out of the 10 inactive patients – chronic patients that died during the pilot programs – was consistent, as it can be seen from the map representation. Also, studying the two age groups, although 58% of the patients are over than 65 years old, it is of particular impression that the 1/3 of them are consistent. Even more impressive is that these 15 consistent elder patients constitute the 71% of the total consistent patients in the dataset. Therefore, the results indicate that chronic patients that are younger than 65 years old are less diligent in taking their measurements and they need much more enhancement and motivation to be engaged with this essential task.

What is more, consistency has an impact on the health condition monitoring as well. When a patient takes their measurements regularly, according to their doctor's instructions, the proposed solution can assess the quality of the measurements and detect dangerous abnormalities of the values. The more values have been recorded, the much more accurate result the performed algorithms will give. Through this way, the patient is able to get early informed about the deterioration of their medical situation and take the necessary measures i.e. inform their doctor timely and be given appropriate recommendations on how to handle their condition. In the same context, user's feedback could be much more valuable and actionable if it was enhanced with visualizations like the "Good and Bad days" one, that can easily be interpreted. When an elderly patient, usually technically illiterate, has the possibility to inspect their daily condition or the progress of their medical situation during the week or the month, based only on the colors of a daily calendar, they will be in a position to understand the risks they are running on their own and act immediately.

Moreover, we identified that the most vulnerable group of patients are men, since men who suffer from chronic diseases outnumber women by two to one and in their vast majority are inconsistent. In the same context, observing the results from the inferential statistical tests, we found out that people aged more than 65 years old, although they are very regular in measurement – taking, the values of their measurements are significantly different, to the worse, compared to the younger patients, a fact that makes them vulnerable. Last but not least, regarding the geolocation of the patients, it turns out that those who live in remote areas of Greece and, as we know, are under medical supervision during the pilot programs, perform much better that the ones that belong to Pilot Athens. This could be interpreted as a very good indicator about how well the healthcare system works in the province, compared to large urban centers, and how much more successfully the personalized medical monitoring of patients is being achieved.

To recapitulate, aiming at improving self - care and enhancing the personalized medical monitoring of chronic patients, the proposed approach can circumvent the challenges of modern electronic health systems. The added value is that it models the behavior of the involved users providing them benefits such as self - monitoring of their medical condition, enhancement of their prosperity and early detection and management of potentially dangerous situations, based on the understanding of their own behavior and performance. At the same time, from the perspective of a telemonitoring system, it can also support the decision-making processes of the patients’ doctors and improve the quality of the provided healthcare.

7.2 Future Perspectives

Taking into consideration the conclusions that were presented in the previous section, along with the pain points that came out from the difficulties that we faced during the analysis procedure, there are many actions could be performed in order to address the challenges of datamining and, as a result, enhance the performance of the proposed solution.

First and foremost, it is undeniable that healthcare professionals are unfamiliar with data mining techniques performed for data analytics purposes. Therefore, it is of vital need that the proposed solution be automated and integrated in a respective healthcare service, in order to be able to offer its added value to the patients. This could be achieved with a cloud-based automated structure that would be able to prevent the medical experts’ usage failures.

Moreover, the biggest problem that we faced was the quality of the initial data, which was the most time-consuming and effort-demanding part of the analysis. During the step of pre-processing, we performed various data harmonization techniques in order to make them in an appropriate format so that they be compatible with the algorithms’ needs. These procedures, including the handling of missing values and the outlier ones along with the data transformations, result in data elimination and deletion, which could possibly signify loss of significant amount of information. As a result, the data will be tangible and easily manageable, but not complete, a fact that could possibly mislead the analysis results. In this context, it is essential to implement better methods of data processing with main focus on the missing value estimation, not elimination, and as far as the outliers are concerned, since they could be highly correlated with forms of rare diseases, it would be worth analyzing them instead of neglecting them.

At the same time, a more effective way of data collection is of imperative need to be used. Not only they are not in a good quality, but also they are not exported properly. As a result, we could not take advantage of the precious information of the patients' physical activity and study it in respect to their medical condition, because the data were in an uninterpretable format. Since everyday walking is an activity that all the doctors suggest to chronic and elder patients, it is a shame that we were not able to take this measure into consideration. However, we were able to notice that there were not enough relevant measurements taken. Consequently, a suggestion for the future is that respective notifications in the object application be added, so that the users be persuaded to enhance their physical activity. In this context, a visualization that would present the evolution of their health in respect to their activity would be of great help.

Furthermore, one of the pain points that we have highlighted several times is patients' inconsistency in keeping a program of measurement – taking, regarding their vital signs. This could be enhanced by adding to the application a personalized, strictly engaging plan with preset notifications and alerts, depending on the needs of the chronic disease the patient suffers from, that could motivate the user to engage more actively. Within this framework, if more interactive and informative visualizations were included in the application, such as the one with the "good and bad days", the patients would be able to control at any time the evolution of their health, and, based on a good feedback, they would be much more motivated to perform the necessary tasks that are detected by their health condition.

Last but not least, a prediction model that would take as input all the findings that we have got so far should be added in the proposed solution, so that the user be timely informed about possible sever implications that could happen to their health in the future, based on their performance so far, early detect them and effectively treat them on time.

ABBREVIATIONS - ACRONYMS

IT	Information Technology
IoT	Internet of Things
NCD	Non-Communicable Diseases
PHR	Personal (Electronic) Health Record
QS	Quantified Self
PPG	Photoplethysmogram
IPF	Idiopathic Pulmonary Fibrosis
AI	Artificial Intelligence

REFERENCES

[1] A. Ala, "Global status report on noncommunicable diseases" World Health Organization Library, Geneva, 2011.

[2] M. Hoogendoorn and B. Funk, Machine Learning for the Quantified Self, Switzerland: Springer, Cham, 2018.

[3] A. Hamper, I. Eigner, N. Wickramasinghe and F. Bodendorf, "Rehabilitation Risk Management: Enabling Data Analytics with Quantified Self and Smart Home Data," in *Health Informatics Meets eHealth*, Amsterdam, IOS Press BV, 2017, pp. 152 - 160.

[4] A. Baerg, "Big Data, Sport, and the Digital Divide," *Journal of Sport and Social Issues,* vol. 41, no. 1, pp. 3 - 20, 2016.

[5] C. Mathers, "WHO methods and data sources for global burden of disease estimates 2000-2016," Department of Information, Evidence and Research WHO, Geneva, 2018.

[6] "The General Meeting of the WHO Global Coordination Mechanism on the Prevention and Control of Noncommunicable Diseases," 2018. [Online]. Available: http://www.who.int/global-coordination-mechanism/events/2018-gcm-general-meeting/en/. [Accessed 16 09 2018].

[7] H. Zhang, K. Liu and W. Kong, "A mobile health solution for chronic disease management at retail pharmacy," in *IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Munich, 2016.

[8] M. Cryer and J. Winkler, "The evolution of the Management of Chronic Illness: World Economic Forum," 2014. [Online]. Available: http://www.ifebp.org/inforequest/0165160.pdf. [Accessed 07 2018].

[9] C. J. L. Murray and e. al., "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010," *THE LANCET,* vol. 380, no. 9859, pp. 2197-2223, 2013.

[10] E. . H. Wagner, B. T. Austin, C. Davis, M. Hindmarsh, J. Schaefer and A. Bonomi, "Improving Chronic Illness Care: Translating Evidence Into Action," *CHRONIC CARE IN AMERICA,* vol. 20, no. 6, 2001.

[11] E. Dermatas, "Embedded construction for the detection, storage and processing of audio biosignals," 10 2016. [Online]. Available: http://nemertes.lis.upatras.gr/jspui/bitstream/10889/10526/1/Epeksergasia%20akoustikwn%20vioshmatwn-revised%2014-6-2017.pdf. [Accessed 09 2018].

[12] M. Swan, "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery," *Big Data,* vol. 1, no. 2, pp. 85-99, 2013.

[13] V. R. Lee, "What's Happening in the "Quantified Self " Movement?," Instructional Technology and Learning Sciences Faculty, Logan UT, 2014.

[14] "The Quantified Self," Quantified Self, 2015. [Online]. Available: http://quantifiedself.com/. [Accessed 09 10 2018].

[15] N. Bermingham-McDonogh, "The Data Science of the Quantified Self," Vrije Universiteit Amsterdam, Amsterdam, 2015.

[16] T. Fawcett, "Mining the Quantified Self: Personal Knowledge Discovery as a Challenge for Data Science," *Big Data,* vol. 3, no. 4, p. 249–266, 2015.

[17] Y. Wang, I. Weber and M. Prasenjit, "Quantified Self Meets Social Media: Sharing of Weight Updates on Twitter," in *6th International Conference on Digital Health*, Montréal, 2016.

[18] M. Honan, "THE NEXT BIG HEALTH APP NEEDS TO DO MORE THAN JUST TRACK OUR NUMBERS," 21 03 2014. [Online]. Available: https://www.wired.com/2014/03/heres-hoping-healthbook-puts-pretty-face-numbers/.

[19] P. R. Sama, Z. J. Eapen, K. P. Weinfurt, B. R. Shah and K. A. Schulman , "An Evaluation of Mobile Health Application Tools," *JMIR Mhealth Uhealth ,* vol. 2, no. 2, pp. 19-25, 2014.

[20] A. Marcengo and A. Rapp, "Visualization of Human Behavior Data: The Quantified Self," in *Innovative Approaches of Data Visualization and Visual Analytics*, Hershey, PA, IGI Global, 2014, pp. 236-265.

[21] M. Ballano Barcena , C. Wueest and H. Lau, "How Safe is your Quantified Self?," Symantec, 2014. [Online]. Available: https://www.symantec.com/connect/blogs/how-safe-your-quantified-self-tracking-monitoring-and-wearable-tech. [Accessed 08 2018].

[22] S. Moyle, "What is mHealth," Ausmed, 03 2015. [Online]. Available: https://www.ausmed.com/articles/what-is-mhealth/. [Accessed 11 2018].

[23] "mHealth New horizons for health through mobile technologies: second global survey on eHealth," WHO Library Cataloguing-in-Publication Data, Geneva, 2011.

[24] E. Wicklund, "Digital Health Tools Help Patients, Doctors Plan for End-of-Life Care," mHealth Intelligence, 11 2016. [Online]. Available: https://mhealthintelligence.com/news/digital-health-tools-help-patients-doctors-plan-for-end-of-life-care. [Accessed 02 2019].

[25] M. Swan, "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0," *Journal of Sensor and Actuator Networks,* vol. 1, no. 3, pp. 217-253, 2012.

[26] A. Georgountzou, A. Menychtas and I. Maglogiannis, "Hypertension Self Management," in *ACM*, Barcelona, 2017.

[27] "Sphygmomanometer," Wikipedia, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Sphygmomanometer. [Accessed 2019].

[28] A. Berger, "Oscillatory Blood Pressure Monitoring Devices," *BMJ ,* vol. 323, no. 919, p. 7318, 2001 .

[29] J. Preece, "What Is a Glucometer and How Does It Work?," Dignifyed, 12 2017. [Online]. Available: https://www.dignifyed.com/glucometer-faqs-review-43.html. [Accessed 04 2018].

[30] J. González, "Spirometer Demo with Freescale Microcontrollers," NXP, 2012.

[31] "Scale," Wiktionary, [Online]. Available: https://el.wiktionary.org/wiki/%CE%B6%CF%85%CE%B3%CE%B1%CF%81%CE%B9%CE%AC. [Accessed 09 2018].

[32] C. Lashkari, "Types of sensors in wearable fitness trackers," News Medical Life Sciences , 23 09 2018. [Online]. Available: https://www.news-medical.net/health/Types-of-sensors-in-wearable-fitness-trackers.aspx. [Accessed 11 2018].

[33] M. Nakka, "Wearable Technology Market Worth USD 71.23 Billion by 2021 - Scalar Market Research," Open PR, 02 2017. [Online]. Available: https://www.openpr.com/news/451708/Wearable-Technology-Market-Worth-USD-71-23-Billion-by-2021-Scalar-Market-Research.html. [Accessed 09 2018].

[34] OECD/ITF, "Big Data and Transport: Understanding and Assessing Options," International Transport Forum, 2015. [Online]. Available: https://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf. [Accessed 07 10 2018].

[35] M. Chen, S. Mao and Y. Liu, "Big Data: A Survey," *Springer,* vol. 19, pp. 171-209, 2014.

[36] "Chauvenet's Criterion," Statistics How To, 05 2016. [Online]. Available: https://www.statisticshowto.datasciencecentral.com/chauvenets-criterion/. [Accessed 10 01 2019].

[37] C. Gorrie, "Outlier Detection," Github, 03 2016. [Online]. Available: http://colingorrie.github.io/outlier-detection.html. [Accessed 03 12 2018].

[38] M. Priya, "IMPUTING THE MISSING VALUES IN IOT USING ESTCP MODEL," *International Journal of Advanced Research in Computer Science,* vol. 8, pp. 532 - 536, 2017.

[39] R. Malarvizhi and A. Selvadoss Thanamani, "K-Nearest Neighbor in Missing Data Imputation," *International Journal of Engineering Research and Development,* vol. 5, no. 1, pp. 05 - 07, 2012.

[40] A. Goel, "How to Manage Noisy Data," Magoosh Data Science Blog, 04 2018. [Online]. Available: https://magoosh.com/data-science/what-is-deep-learning-ai/. [Accessed 02 2019].

[41] M. Pathak, "Introduction to t-SNE," DataCamp, 13 09 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/introduction-t-sne. [Accessed 02 2019].

[42] R. Silipo, "Seven Techniques for Data Dimensionality Reduction," KDnuggets, 2015. [Online]. Available: https://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html. [Accessed 02 2019].

[43] A. Savitzky and M. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry,* vol. 36, no. 8, p. 1627–1639, 1964.

[44] H. Lohninger, "Savitzky-Golay Filter," Fundamentals of Statistics, [Online]. Available: http://www.statistics4u.com/fundstat_eng/cc_filter_savgolay.html#.

[45] "What are the best normalization techniques in data mining?," Quora, 09 2013. [Online]. Available: https://www.quora.com/What-are-the-best-normalization-techniques-in-data-mining. [Accessed 12 2018].

[46] S. Raschka, "About Feature Scaling and Normalization – and the effect of standardization for machine learning algorithms," Sebastian Raschka, 07 2014. [Online]. Available: https://sebastianraschka.com/Articles/2014_about_feature_scaling.html#about-standardization. [Accessed 11 2018].

[47] C. Fryar, Q. Gu and C. Ogden, "Anthropometric Reference Dta for Children and Adults: United States 2017 - 2010," *Vital Health Stat,* vol. 11, p. 252, 2012.

[48] "Analyzing Quantitative Research," Center of Innovation in Research and Teaching , [Online]. Available: https://cirt.gcu.edu/research/developmentresources/research_ready/quantresearch/analyze_data. [Accessed 09 2018].

[49] E. Taylor-Powell, "Analyzing Quantitative Data," University of Wisconsin, 2019.

[50] P. Tobias and C. Croarkin , "e-Handbook of Statistical Methods," NIST/ SEMATECH, 2013. [Online]. Available: https://www.itl.nist.gov/div898/handbook/. [Accessed 2019].

[51] "Welch's t-test," Python for Data Science, 2019. [Online]. Available: https://pythonfordatascience.org/welch-t-test-python-pandas/. [Accessed 2019].

[52] G. Drakos, "Handling Missing Values in Machine Learning," Towards Data Science, 08 2018. [Online]. Available: https://towardsdatascience.com/handling-missing-values-in-machine-learning-part-2-222154b4b58e. [Accessed 10 2018].

[53] V. Konovalov, "How can I choose the best K in KNN (K nearest neighbour) classification?," Quora, 01 2018. [Online]. Available: https://www.quora.com/How-can-I-choose-the-best-K-in-KNN-K-nearest-neighbour-classification. [Accessed 10 2018].

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchessnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825 - 2830 , 2011.

[55] "Agglomerative Hierarchical Clustering," Datanovia, 2018. [Online]. Available: https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/. [Accessed 2019].

[56] "Complete-linkage clustering," Wikipedia, 10 2018. [Online]. Available: https://en.wikipedia.org/wiki/Complete-linkage_clustering. [Accessed 11 2018].

[57] O. &. R. K. T. Yim, "Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data," *The Quantitative Methods for Psychology ,* vol. 11, no. 1, pp. 8 - 21, 2015.

[58] "Scipy Cluster Hierarchy Linkage," SciPy.org, 2016. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage. [Accessed 2018].

[59] A. Georgountzou and I. Maglogiannis, "HeartAround: Advanced System for Monitoring and Support of Chronic Patients," in *11th International Forum of the 23rd Scientific Congress for Hellenic Medical Students*, Larisa, Greece, 2017.

[60] "Are mean normalization and feature scaling needed for k-means clustering," StackExchange, [Online]. Available: https://stats.stackexchange.com/questions/21222/are-mean-normalization-and-feature-scaling-needed-for-k-means-clustering. [Accessed 02 02 2019].

[61] S. O'Grady, "The RedMonk Programming Language Rankings: January 2013," RedMonk, 02 2013. [Online]. Available: https://redmonk.com/sogrady/2013/02/28/language-rankings-1-13/. [Accessed 01 2019].

[62] "HTML," Wikipedia, [Online]. Available: https://simple.wikipedia.org/wiki/HTML. [Accessed 03 02 2019].

[63] T. E. Oliphant, Guide to Numpy, USA : ACM, 2015.

[64] W. McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, 2010.

[65] "SciPy library," SciPy.org, 2019. [Online]. Available: https://www.scipy.org/scipylib/index.html. [Accessed 2019].

[66] "Hierarchical clustering," ScyPi.org, 11 05 2014. [Online]. Available: https://docs.scipy.org/doc/scipy-0.14.0/reference/cluster.hierarchy.html. [Accessed 12 2018].

[67] M. Farber, "Missing value imputation in python using KNN," Stackoverflow, 2017. [Online]. Available: https://stackoverflow.com/questions/45321406/missing-value-imputation-in-python-using-knn. [Accessed 2018].

[68] H. John, D. Darren, F. Eric and D. Michael, "Matplotlib," 2018. [Online]. Available: https://matplotlib.org/. [Accessed 2018].

[69] M. A. Barrett, O. Humblet, R. A. Hiatt and N. E. Adler, "BIG DATA AND DISEASE PREVENTION: From Quantified Self to Quantified Communities," *Big Data,* vol. 1, no. 3, pp. 168-175, 2013.

[70] G. Wolf, "The data-driven life," NewYorkTimes, [Online]. Available: https://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html. [Accessed 12 09 2018].

[71] "Annual Business Plan 2009/10 - 2011/12," Waterloo Wellington LHIN , Ontario, 2009.

[72] S. Miller, "NIST maps out IoT security standards," GCN, 02 2018. [Online]. Available: NIST maps out IoT security standards. [Accessed 05 2018].

[73] U. Malik, "Hierarchical Clustering with Python and Scikit-Learn," Stack Abuse, 07 2018. [Online]. Available: https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/.

[74] "Big Data Analytics - K - Means Clustering," TutorialsPoint , 2018. [Online]. Available: https://www.tutorialspoint.com/big_data_analytics/k_means_clustering.htm. [Accessed 2018].

[75] Y. &. F. D. Choe, "The Quantified Traveler: Implications for Smart Tourism Development," in *Analytics in Smart Tourism Design: Concepts and Methods*, Florida , Springer, 2017, pp. 65-67.

[76] B. O'Brien, "The Internet Of Medicine Is Just What The Doctor Ordered," Tech Crunch, 2016. [Online]. Available: https://techcrunch.com/2016/02/16/the-internet-of-medicine-is-just-what-the-doctor-ordered/?guccounter=1. [Accessed 2018].

[77] William M.K. Trochim, "Social Research Methods," Web Center for Social Research Methods, 2006.