

# **Clustering of physical activity patterns in COPD patients**

## **Bachelor's Thesis in Medical Engineering**

submitted  
by

Kevin Rättsch

born March 31, 1992 in Celle

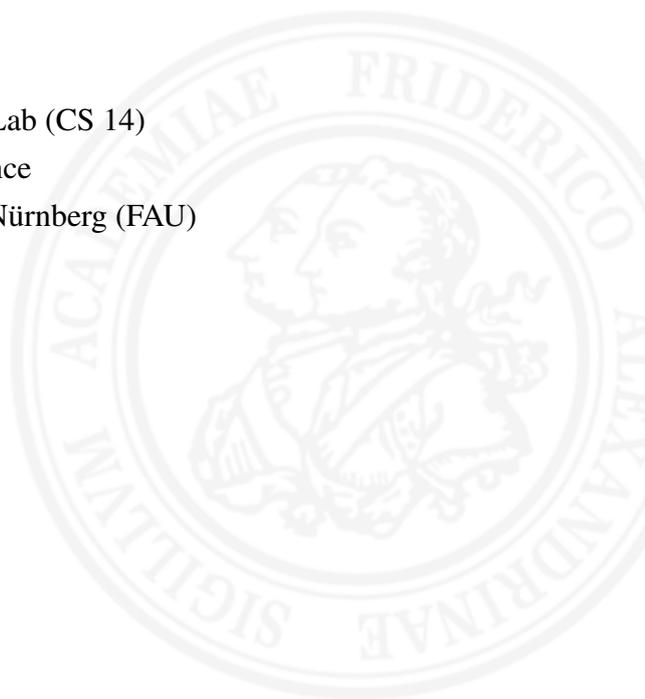
Written at

Machine Learning and Data Analytics Lab (CS 14)  
Department of Computer Science  
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Advisors: Martin Ullrich  
Nils Roth  
Dr. Wolfgang Geidl  
Prof. Dr.-med. Jochen Klucken  
Prof. Dr. Björn Eskofier

Started: 29.10.2018

Finished: 29.03.2019





Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 29.03.2019



## Abstract

For patients with chronic obstructive pulmonary disease (COPD) physical activity (PA) is associated with reduced lung function decline and an improvement of the overall health status. Therefore, the goal of pulmonary rehabilitation (PR) is to increase PA in COPD patients, but often it does not have a sustainable impact on the patients behaviour. To understand why, a better understanding of the PA behaviour in COPD patients is needed. In this thesis, a cluster analysis was applied to analyze daily PA of COPD patients and to distinguish daily activity profiles. In a data-driven approach 92 features were defined as well as seven features in an expert-driven approach. For the cluster analysis a hierarchical clustering algorithm and the k-means algorithm were applied and the silhouette method was used as an internal evaluation criterion. As an external evaluation criterion, an analysis of variance (ANOVA) was applied to additional clinical parameters, that were not used for the clustering algorithm. The silhouette method identified the best cluster separation for the k-means algorithm within the expert-driven approach. Four clusters could be identified which show all significant differences in clinical parameters. These findings could contribute to a better understanding of PA in COPD patients and therefore could be used to develop individualized rehabilitation strategies.

## Übersicht

Bei Patienten mit chronisch obstruktiver Lungenerkrankung (COPD) führt körperliche Aktivität (KA) zu einer verringerten Lungenfunktionsstörungen und zu einer Verbesserung des allgemeinen Gesundheitszustandes. Ziel der pulmonalen Rehabilitation ist es daher, die KA bei COPD-Patienten zu erhöhen, oft hat diese jedoch keinen nachhaltigen Einfluss auf das Verhalten des Patienten. Aus diesem Grund ist ein besseres Verständnis des Aktivitätsverhaltens von COPD Patienten erforderlich. In dieser Arbeit wurde die tägliche KA von COPD Patienten mit Hilfe einer Clusteranalyse analysiert, um tägliche Aktivitätsprofile zu identifizieren. In einem datengetriebenen Ansatz wurden 92 Merkmal definiert sowie sieben Merkmale in einem expertengetriebenen Ansatz. Für die Clusteranalyse wurden ein hierarchischer Algorithmus und der k-means-Algorithmus eingesetzt. Die Silhouetten-Methode wurde als internes Bewertungskriterium verwendet. Als externes Bewertungskriterium wurde eine Varianzanalyse (ANOVA) auf weitere klinische Parameter angewendet, die nicht für das Clustering verwendet wurden. Die Silhouettenmethode identifizierte die beste Clustertrennung für den k-means-Algorithmus innerhalb des expertengetriebenen Ansatzes. Es konnten vier Cluster identifiziert werden, die alle signifikanten Unterschiede in den klinischen Parametern aufweisen. Diese Ergebnisse könnten zu einem besseren Verständnis von KA bei COPD Patienten beitragen und somit zur Entwicklung individueller Rehabilitationsstrategien genutzt werden.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Role of Pulmonary Rehabilitation . . . . .	2
1.2	Monitor Physical Activity in COPD Patients . . . . .	3
1.3	Analyzing Physical Activity in a Meaningful Way . . . . .	3
1.4	Clustering of Physical Activity . . . . .	4
1.5	Purpose of this Thesis . . . . .	5
1.6	Research Questions . . . . .	5
1.7	Outline . . . . .	5
<b>2</b>	<b>Fundamentals</b>	<b>7</b>
2.1	Chronic Obstructive Pulmonary Disease . . . . .	7
2.1.1	Prevalence . . . . .	7
2.1.2	Disease Development and Progression . . . . .	8
2.1.3	Diagnosis and Symptoms . . . . .	8
2.1.4	Clinical Assessment . . . . .	9
2.1.5	Pulmonary Rehabilitation . . . . .	10
2.2	Actigraphy . . . . .	12
<b>3</b>	<b>Methods</b>	<b>13</b>
3.1	Data Source . . . . .	13
3.2	Data Pre-Processing . . . . .	14
3.3	Clustering of PA Barcodes . . . . .	15
3.3.1	Feature Extraction - Data-Driven Approach . . . . .	16
3.3.2	Feature Extraction - Expert-Driven Approach . . . . .	16
3.3.3	Dimensionality Reduction . . . . .	17
3.4	Clustering . . . . .	18

3.4.1	Hierarchical Clustering . . . . .	18
3.4.2	Partitional Clustering . . . . .	19
3.5	Clustering Evaluation . . . . .	20
3.5.1	Silhouette Method . . . . .	20
3.5.2	External Evaluation . . . . .	21
<b>4</b>	<b>Results</b>	<b>23</b>
4.1	Data-Driven Approach . . . . .	23
4.1.1	Hierarchical Clustering . . . . .	25
4.1.2	k-Means . . . . .	29
4.2	Expert-Driven Approach . . . . .	33
4.2.1	Hierarchical Clustering . . . . .	35
4.2.2	k-Means . . . . .	39
<b>5</b>	<b>Discussion</b>	<b>45</b>
5.1	Data Pre-Processing . . . . .	45
5.2	Data-Driven vs. Expert-Driven Approach . . . . .	46
5.3	Hierarchical vs. Partitional Clustering . . . . .	47
5.4	Interpretation of the Clustering Results . . . . .	48
5.5	Strengths and Limitations . . . . .	50
<b>6</b>	<b>Conclusion and Outlook</b>	<b>51</b>
<b>A</b>	<b>Glossar</b>	<b>53</b>
<b>B</b>	<b>Patents</b>	<b>55</b>
B.1	US8337431B2 . . . . .	55
B.2	US8543185B2 . . . . .	56
B.3	US9737261B2 . . . . .	56
<b>C</b>	<b>COPD Assessment Test</b>	<b>57</b>
<b>D</b>	<b>List of Features (Data-Driven Approach)</b>	<b>59</b>
	<b>List of Figures</b>	<b>63</b>
	<b>List of Tables</b>	<b>65</b>

*CONTENTS*

ix

**Bibliography**

**67**



# Chapter 1

## Introduction

Regular physical activity (PA) has many well established health benefits and reduces the risk of many chronic diseases [War06]. Especially for patients with chronic obstructive pulmonary disease (COPD) PA is associated with reduced lung function decline and improves the overall health status [GA07, Spr13]. PA in patients with COPD is mostly described as a total amount of activity per day or as an average of multiple day measurements [Wat09, Pit08, Hil12, Ega12]. The review of Byrom and Rowe summarized the used methodologies and measured results derived from the use of accelerometers to measure free-living activity in patients with COPD [Byr16]. They found that over 50% of the included studies were only reporting the total activity per day or per hour. Contrary to this, Bussmann et al. described PA as a multidimensional construct and suggested that PA should be described and analyzed in more detail to gain more insights about the effects of PA on patients with chronic diseases [Bus13]. Especially for the evaluation of pulmonary rehabilitation (PR) strategies for COPD patients, a more detailed understanding of PA can provide valuable insights into the effects of PR.

## 1.1 The Role of Pulmonary Rehabilitation

PR is defined as “a comprehensive intervention based on thorough patient assessment followed by patient-tailored therapies that include, but are not limited to, exercise training, education, self-management intervention aiming at behavior change, designed to improve the physical and psychological condition of people with chronic respiratory disease and to promote the long-term adherence to health-enhancing behaviors” [Spr13].

The benefits for COPD patients from PR are considerable. In the official American Thoracic Society & European Respiratory Society Policy Statement the authors stated that the provision of PR as an evidence-based and standardized component of the overall integrated patient care with COPD should be increased [Roc15]. PR would not only improve the physical and emotional health and quality of life of the patient, but will also improve the quality of patient care and has the potential to significantly reduce health care costs over time. Inactive patients with COPD are reported to have worse exercise capacity, more dyspnea and a worse functional status which can lead patients into a vicious cycle of increased dyspnea, declining lung function and mortality [Pit06b]. Increased PA of COPD patients is associated with improved health outcomes including reductions in hospital admissions and respiratory mortality [GA06]. PR has also been shown to be the most effective therapeutic strategy to improve shortness of breath, health status and exercise tolerance [McC15]. It is appropriate for most patients with COPD and improved functional exercise capacity and health related quality of life have been demonstrated across all grades of COPD severity [Sah16].

Despite the improvements made during PR, most patients return to pre-rehabilitation levels of physical endurance within 6-24 months of program discharge [Fog99, Tro00]. Recent studies also show that PR programs often fail to sustainably increase physical activity [Bus14, CN12, Spr15, GS14]. A closer look at the patient’s daily PA pattern could contribute to a better understanding of the effects of PR and would help to develop effective rehabilitation strategies to improve the patients PA behaviour.

Geidl et al. investigated, in a randomized controlled trial, the additional effect of a pedometer-based behavior-change intervention during inpatient PR on objectively measured PA [Gei17]. They measured PA before the rehabilitation as well as 6 weeks and 6 months after the rehabilitation. While Geidl et al. focused their research on the PA patterns of a week, the same data can also be analyzed in the scope of a single day to get a better understanding of the determinants of PA in patients with COPD.

## 1.2 Monitor Physical Activity in COPD Patients

Several studies have been conducted to monitor the PA of patients, including direct observation, self-report questionnaires and activity accelerometry monitoring. While direct observation is impractical over long periods of time, self-report questionnaires failed to accurately assess PA [Pit06a]. Among activity monitoring methods, accelerometry has been shown to be more accurate than pedometry, especially in COPD patients [Ste03, Pit05]. The review of Byrom and Rowe (2016) summarized 76 studies published between 1999 and 2014 that were using activity monitoring devices to track PA in COPD patients [Byr16]. Therefore, there should be no technical barriers for more detailed analysis of PA.

Furthermore, there are several patented systems that are related to PA monitoring (see Appendix B). In Appendix B.1 an implantable medical device that determines when a patient is attempting to sleep is proposed by Heruth and Miesel. During the day, the device periodically determines the patient's activity levels and it also determines values for a variety of metrics that indicated the quality of a patient's sleep. This information could be used to evaluate the effectiveness of a therapy. The invention that was patented by Yuen et al. (Appendix B.2) is a portable activity monitoring system, that can detect the activity of the user and generate data which is representative of the user's PA. Whereas the first presented patent is directly related to health care, the idea proposed by Yuen et al. is focusing on a portable system that can reliably monitor the PA of the user. Coza et al. (Appendix B.3) have enhanced the idea of a portable PA monitoring system and invented a sensor garment to monitor an individual engaged in an athletic activity. The sensor module includes a single-purpose sensor configured to sense a single characteristic, and a radio antenna configured to transmit data generated by the single-purpose sensor. The patent review reveals that PA monitoring is used in various applications, which are more or less related to health care and sports.

## 1.3 Analyzing Physical Activity in a Meaningful Way

Studies with different populations have shown that more detailed insights can be generated by analysis of different PA parameters than by comparing the total amount of PA in a day. Bussmann et al. stressed the importance of those alternative constructs and parameters and concluded that the challenge for the future will be to determine which parameters are most relevant, valid and responsive [Bus13]. Since the same amount of daily PA can be achieved by many short bouts interspersed throughout the day or from few long bouts, there needs to be a more precise measurement of PA. For example, 60 minutes moderate PA during a day can be composed of 30

bouts of two minutes, or two bouts of 30 minutes. The physiological effect of those two examples will be considerably different. The importance of bout length and the bout distribution is well described in the reviewed literature. A study by Healy et al. showed that the total amount of sedentary time should be decreased to reduce the risk for cardiovascular diseases, but also that interrupt sedentary time can have a beneficial impact [Hea11]. This finding can be supported by Chastin et al. who showed that patients with Parkinson's disease compared to healthy subjects did not differ in the total amount of sedentary time, but did significantly differ in the distribution of sedentary time during the day [Cha10]. In studies on PA behavior, the total measurement period is often aggregated to an average value and possible effects are averaged out. For example, a study by Rochester et al. compared volume measures such as amount of time walking, amount of time standing, and number of walk periods in subjects with Parkinson disease vs. healthy control subjects [Roc06]. Whole-day analysis showed no significant difference, while the same data expressed on a hourly basis did. This means that the relevance of the outcome depends not only on relevant measures but also on the time window of analysis.

Furthermore, Paraschiv-Ionescu et al. suggested the concept of 'barcoding' for analyzing PA in chronic pain patients [PI12]. Combining different features of PA (type, intensity, duration) to define various PA states allowed the creation of a temporal sequence of different states. The temporal sequence then was visualized as a 'barcode'. With this approach, daily activity information can be contained in the temporal structure of the barcode. The research group found, that significant information about pain-related functional limitations can be captured in the structural complexity of PA barcodes.

## 1.4 Clustering of Physical Activity

In order to get even more insights, cluster analysis can be a useful tool to identify subgroups of patients with distinct PA characteristics. The first study that applied cluster analysis to accelerometer data was conducted by Lee et al. [Lee13]. The authors were able to identify two clusters of subjects in middle-aged Chinese adults, one more active than the other and could correlate the clusters to health characteristics like the body fat percentage. Mesquita et al. applied cluster analysis to a total of 1001 patients with COPD [Mes17]. Based on PA measures and hourly patterns they could identify five clusters and characterize them based on demographics, lung function and clinical data. These detailed analyses could then lead to new insights regarding subgroups of patients with COPD with specific physical activity patterns, which may be used in further investigations and intervention strategies.

## 1.5 Purpose of this Thesis

The literature review shows, that a greater amount of studies only reported the total amount of activity per day or per hour, but did not differentiate the behaviour that leads to the amount of PA. This can lead to relevant differences in PA not being detected. Especially in context of PR a closer look at the patient's PA behaviour can contribute to a better understanding of the effectiveness of rehabilitation strategies.

Thus, the purpose of this thesis was to analyze the data, provided by the STAR study (Stay Active after Rehabilitation) [Gei17], with a high time resolution in order to identify typical PA patterns throughout a day and to identify subgroups of patients with different PA behaviour and health characteristics. Different approaches for the feature extraction and different clustering algorithms were applied to evaluate, which approaches and algorithms are most effective to cluster PA in COPD patients.

## 1.6 Research Questions

The following research questions were developed for this thesis:

- Is it possible to identify clusters with different PA behaviour based on accelerometer data?
- Can those clusters objectively distinguish daily profiles of PA?
- Which clustering approach can provide the best differentiation?

## 1.7 Outline

The remainder of the thesis is organized as follows: Chapter 2 provides basic information on the prevalence, diagnosis, assessment and rehabilitation of COPD and on the functionality of the ActiGraph monitor. In Chapter 3 the methodology is described. The results of the cluster analysis is presented in Chapter 4 and discussed in Chapter 5. Finally, Chapter 6 provides a conclusion of the thesis with open questions that should be answered in future studies.



# Chapter 2

## Fundamentals

### 2.1 Chronic Obstructive Pulmonary Disease

As described in the Executive Summary of the Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease, COPD is a common, preventable and treatable disease, that is characterized by persistent respiratory symptoms and airflow limitation caused by airway and/or alveolar abnormalities [Vog17]. These abnormalities are usually caused by a significant exposure to noxious particles or gases and the most common symptoms include dyspnea, cough and sputum production.

#### 2.1.1 Prevalence

COPD is one of the leading causes of morbidity and mortality worldwide [Loz12]. The prevalence, morbidity and mortality vary across countries and they are the result of a complex interplay of long-term cumulative exposure to noxious particles or gases as well as a variety of host factors including genetics, airway hyper-responsiveness and poor lung growth during childhood, which can predispose individuals to develop COPD [Lan15, Ste07, Tas92]. However, determining the existing COPD prevalence is complex because the data vary widely due to differences in survey methods, diagnostic criteria and analytic approaches [Mat06]. According to a systematic analysis for the global burden of diseases by Murray et al., approximately 329 million people (around 4,77% of the population) were affected by COPD in 2010 [Mur12]. Besides, a systematic review and meta-analysis by Halbert et al., including studies from 28 countries between 1990 and 2004, provided evidence that the prevalence of COPD is significantly higher in smokers and ex-smokers, compared to non-smokers as well as in those  $\geq 40$  years of age compared to those  $< 40$ , and also

in men compared to women [Hal06]. Furthermore the prevalence of COPD is likely to be directly related to the prevalence of tobacco smoking and other environmental exposures such as biomass fuel exposure and air pollution [Eis10, Sal09].

### 2.1.2 Disease Development and Progression

COPD results from a complex interaction between genes and the environment. Although cigarette smoking is the leading environmental risk factor for COPD, even within the group of heavy smokers, less than 50% develop COPD during their lifetime [Ren06]. Genetics may modify the risk of COPD in smokers and there may also be other risk factors involved.

For instance, a generic risk factor is a severe hereditary deficiency of alpha-1 antitrypsin (AATD), a major circulating inhibitor of serine proteases [Sto05]. Furthermore, a significant family-related risk of airflow limitation has been observed in people who smoke and are siblings of patients with severe COPD [McC01]. Additionally, several genome-wide associated studies have linked genetic loci with COPD. But it remains uncertain if those genes are directly responsible for COPD or are merely markers of casual genes [Cho10, Pil09, SA11, Rep10, Cho14]. Age is also often listed as a risk factor for COPD, but it is unclear whether healthy aging as such lead to COPD or if age reflects the sum of cumulative exposures throughout life [Mer15].

Incidents occurring during pregnancy and birth as well as exposure during childhood and adolescence can affect lung growth [Bar91, Tod93]. A study by Lange et al. evaluated three different longitudinal cohorts and found that approximately 50% of patients developed COPD due to accelerated decline of lung function over time, while the other 50% developed COPD due to abnormal lung growth and development with a normal decline in lung function over time [Lan15].

### 2.1.3 Diagnosis and Symptoms

COPD should be considered in any patient who has dyspnea, chronic cough or sputum production and/or a history of exposure to risk factors [Yaw09]. But to make a diagnosis in a clinical context, a spirometry is required. Spirometry measures the volume of air forcibly exhaled from the point of maximal inspiration (forced vital capacity, FVC) and the volume of air that can forcibly be blown out in one second after full inspiration (forced expiratory volume in the first second, FEV<sub>1</sub>). The calculated ratio of FEV<sub>1</sub>/FVC < 0.70 confirms the presence of airflow limitations and thus the presence of COPD in patients with appropriate symptoms [Bui07]. Chronic and progressive dyspnea is the most characteristic symptom of COPD next to less characteristic symptoms like cough or sputum production and is a major cause of the disability and anxiety that is associated

with the disease [Mir14].

Chronic cough on the other hand is often the first symptom of COPD, but is frequently ignored by patients as an expected consequence of smoking or environmental exposures. With coughing, COPD patients commonly raise small quantities of tenacious sputum. Sputum production as a symptom is often difficult to evaluate because patients may swallow sputum rather than expectorate it [Vog17]. Also to mention are the concomitant chronic diseases that occur frequently in COPD patients, including cardiovascular disease, skeletal muscle dysfunction, metabolic syndrome, osteoporosis, depression, anxiety, and lung cancer [Vog17].

### 2.1.4 Clinical Assessment

In order to guide the therapy, the goals of the assessment are to determine the level of airflow limitation, to define its impact on the health status of the patient and to identify the risk of future events such as exacerbations, hospital admissions or death. To achieve these goals, the COPD assessment must address the following aspects of the disease separately [Vog17]:

- The presence and severity of the spirometric abnormality
- Current nature and magnitude of the patient's symptoms
- Exacerbation history and future risk
- Presence of comorbidities

The FEV<sub>1</sub> cutoff-point that are used for classification of airflow limitation severity in COPD as well as the four corresponding GOLD statuses can be found in Table 2.1. However, Jones et al. showed, that there is only a weak correlation between FEV<sub>1</sub>, symptoms and impairment of a patient's health status [Jon09a]. Therefore, a formal and symptomatic assessment is required.

GOLD Status	Severity	FEV <sub>1</sub>
1	Mild	≥ 80 %
2	Moderate	50 - 80 %
3	Severe	30 - 50 %
4	Very Severe	≤ 30 %

Table 2.1: Classification of airflow limitation severity in COPD in patients with FEV<sub>1</sub>/FVC < 70%

The widely used COPD Assessment Test (CAT<sup>TM</sup>) is a short, eight-item questionnaire, which measures the health status impairment in COPD patients. The CAT<sup>TM</sup> is considered as valid, reliable and standardized measure of COPD health status with worldwide relevance [Jon09b]. One way to interpret the CAT<sup>TM</sup> result is by the COPD ladder of severity (Tab. 2.2) as proposed by Jonas et al. [Jon11]. Representative items for each 5-point step along the CAT<sup>TM</sup> are listed in ascending order of severity. At each level it is likely that the patient will also have experienced the development of many of the health affects associated with the milder steps up to their current severity. As a test for the functional status of the patient, the 6-min walk test (6MWT) can be used. The distance walked in the 6MWT can help to predict the mortality in patients with COPD [Cas08].

### **2.1.5 Pulmonary Rehabilitation**

According to the report of the Global Initiative for Chronic Obstructive Lung Disease, PR is the most effective therapeutic intervention for reducing dyspnea and improving physical performance as well as the quality of life in COPD patients [Vog17]. PR is an intervention that is usually delivered by a multidisciplinary rehabilitation team, in which a comprehensive assessment forms the basis of an individual rehabilitation program. Obligatory components of such a program include exercise training and breathing retraining, patient education on inhalation techniques and nutritional advice, treatment optimization and exacerbation management [Glo18]. The effectiveness of PR is supported by highest-level evidence [Rie07, McC15]. However, the effects on the PA behaviour caused by the PR has also been questioned by several studies, because it fails to substantially enhance PA long-term [Bus14, CN12, Spr15, GS14]. The exercise therapy is a personalized combination of endurance training, strength training and respiratory muscle training and can also include neuromuscular electrical stimulation or whole body vibration training. The primary objective of educating the patients during PR is to cause patients change their behavior. Positive examples for behavior changes in COPD patients are an improved adherence to medication, a continuation of the exercise and dietary modifications, an increased PA, smoking cessation and the use of energy saving strategies during activities of daily life [Glo18].

CAT score	Descriptions
40	Cannot move far from bed or chair Have become frail or an invalid Cannot do housework
35	Cannot take bath/shower or takes a long time Breathless walking around the home Chest trouble has become a nuisance to friends/relatives
30	Everything seems too much of an effort No good days in the week Stops patient doing most of what they want to do
25	Feel that not in control of chest problem Cough/breathing disturbs sleep Get afraid or panic when cannot get breath
20	Wheeze worse in the morning Breathless on bending over Wheezing attacks on most days
15	Cough several days a week Breathlessness on most days Housework takes a long time or have to take rests
10	Usually cannot play sports or games Gets exhausted easily Walk slower than other people or stop for rests
5	Breathlessness stops patient doing one or two things Chest condition causes a few problems Breathless walking up hills

Table 2.2: COPD ladder of poor health

## 2.2 Actigraphy

A widely used activity monitor to assess free-living PA is the ActiGraph (Pensacola, Florida, USA) [Wel02, Sas11]. The latest ActiGraph generation GT3X was developed in 2009 using a triaxial capacitive microelectromechanical system sensor with a full-scale range of  $\pm 3$  Gs [Joh12]. The small and light GT3X accelerometer (27 g; 4.6 cm x 3.3 cm x 1.5 cm) is capable of recording accelerations in three axes (vertical, antero-posterior and medio-lateral) and can be worn on the wrist or hip. For an accurate assessment of sedentary and non-sedentary behaviour a hip-worn sensor is recommended [Byr16].



Figure 2.1: ActiGraph wGT3x-BT [Act09]

The main output of the ActiGraph GT3X (Fig. 2.1) are activity counts that are generated through several processing steps of the original raw acceleration signal [Try96]. Activity counts can be used to determine PA intensity as first shown by Freedson et al. [Fre98]. Several studies validated the accelerometry-based PA monitors and confirmed their correlation to different levels of PA intensity and their energy expenditure [Pat93, Swa00, Wel00]. Santos-Lozano et al. also confirmed the reliability of the ActiGraph GT3X as an accurate tool to estimate free-living physical activity [SL12].

# Chapter 3

## Methods

The group of subjects and the signal processing steps for the feature extraction and cluster analysis will be presented in the following sections. All described methods and calculations were implemented in python (version 3.6).

### 3.1 Data Source

The investigations in this thesis are based on the data provided by the STAR study (Stay Active after Rehabilitation) by Geidl et al. [Gei17]. The study took place within the German rehabilitation system which typically provides an inpatient rehabilitation for a duration of three weeks. Eligible to participate were patients with COPD who are enrolled for PR with the rehabilitation clinic Bad Reichenhall. The goal of the STAR study was to investigate the additional effect of a pedometer-based behavior-change intervention during inpatient pulmonary rehabilitation on objectively measured PA. The physical activity was measured by using a hip-worn physical activity monitor (ActiGraph wGT3X-BT - Pensacola, Florida, USA) for a period of at least 7 days, firstly two weeks before rehabilitation as well as 6 weeks and 6 months after rehabilitation. For the scope of this thesis, only the data from the first period was included. In total, 418 patients participated in the study. Among those, 75 participants had to be excluded from the analysis due to refuted COPD status (n=62), retrospective withdrawal of consent to data use (n=1) or non-participation in the first measurement period (n=12). This led to an overall sample of 343 participants (234 men and 109 women). The participants were asked to wear the ActiGraph during waking hours and outside of any water-based activities. On average, patients wore the ActiGraph for 14 ( $\pm 3$ , 8-29) days. In total, 5083 day measurements were recorded during the first measurement period. General characteristics of the participants can be found in Table 3.1.

General Characteristics	n	Mean	SD	Range
Weight (kg)	331	81	$\pm 22$	35-177
Height (cm)	329	171	$\pm 9$	148-195
Age (years)	343	58	$\pm 6$	43-85
FEV <sub>1</sub> (%)	328	53.5	$\pm 18.3$	15.5-102.5
CAT <sup>TM</sup>	236	23.5	$\pm 6.7$	2-39
6MWT (m)	317	448.5	$\pm 103.0$	120-715

Table 3.1: General characteristics of the COPD patients

## 3.2 Data Pre-Processing

For further data processing only valid days were included. A day was considered valid, when the participant wore the device for ten or more waking hours [Byr16]. Non-wear time was defined by at least 60 minutes of zero counts of which up to two minutes may be within the 0 - 100 count range [Gei17]. Days with at least 60 minutes of non-wear time were also excluded from the further data processing. This results in a total of 2255 valid days. For the scope of this thesis, only the first ten hours of a valid day were used to improve the comparability.

The ActiGraph assessed acceleration by transforming the raw signal into cumulated activity counts attributable to different intensity categories. For the purpose of this thesis, the activity counts were cumulated to 120 seconds intervals. In order to interpret the activity counts, the cut-off points introduced by Freedson et al. [Fre98] were used to characterize activity by intensity. The classification was further refined by the sedentary cut-off definition by Evenson et al. [Eve15]. Therefore, Sedentary behaviour was defined as  $< 100$  counts/minute, light intensity as 101-1951 counts/minute, moderate intensity as 1952-5724 counts/minute, vigorous intensity as 5725-9498 counts/minute and very vigorous intensity as  $> 9498$  counts/minute.

To reduce the complexity of the barcode in the intensity dimension, the higher intensity categories were combined, which results into the following categories: Inactive was defined as  $< 100$  counts/minute, Light Active as 101-1951 counts/minute and High Active as  $> 1951$  counts/minute. By this definition every 120 seconds time interval of the day was tagged with an intensity label. A numerical state was assigned to each intensity label in order to encode the activity behaviour during the observation period into a numerical sequence. The basic idea behind this encoding is to combine different dimensions of physical activity as duration, intensity, frequency and temporal patterns. The sequence can then be analyzed to provide physical activity

metrics and can be represented as a colour barcode to provide global illustrative visual information (Fig. 3.1) [PI12].

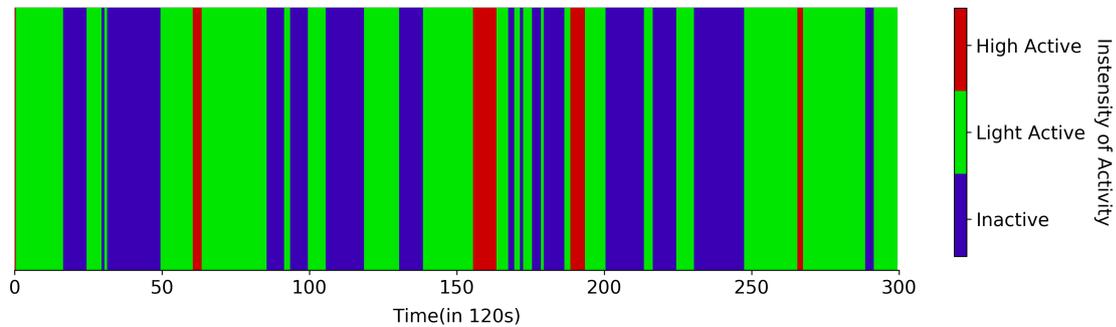


Figure 3.1: Example of a barcode with a resolution of 120s

Additionally a moving median filter with a window-size of 15 data points was applied to reduce spiky noise [Jus81]. The moving median filter processed the input data such that each data point in the output data is the median of a window of points centered on the corresponding point in the input data.

### 3.3 Clustering of PA Barcodes

Each data point of the PA barcode represents a PA state calculated for time windows of 120 seconds. For an analysis period of ten hours this results into 300 data points per barcode. In order to perform the cluster analysis, features were extracted to describe the barcode. Two different approaches for the feature extraction were applied: In a pure data-driven approach, 92 features were extracted and in an expert-driven approach, seven features were selected by the research group of the STAR Study [Gei17]. Furthermore, principal component analysis (PCA) was used for dimensionality reduction. Finally, a partitioning and a hierarchical clustering method were used to separate the barcodes into clusters with distinct characteristics. The silhouette method was used to evaluate the results as well as additional clinical parameters, that were not used for the clustering algorithm. In the following sections the steps of the pipeline will be explained in more detail.

### 3.3.1 Feature Extraction - Data-Driven Approach

In this approach, multiple features were extracted in order to describe the complexity and structural patterns of the barcode. Basic features such as time per intensity and longest time in a sedentary state were calculated. To describe the complexity of the barcode and also retrieve information from the variety, temporal changes and duration of different activity levels, the information entropy was used to calculate the overall entropy of the barcode. In Equation 3.1, the entropy  $H(\mathbf{x})$  is calculated by the negativ sum of the probability mass function  $p(x_i)$  multiplied by the logarithm of  $p(x_i)$ .

$$H(\mathbf{x}) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (3.1)$$

In addition, the entropy was calculated for two modified, binary barcode variations. The first variation only distinguished between inactive and active while the second variation only distinguished between high active and not high active. The entropy score takes large (small) values when there are many (few) changes of the activity level in the barcode [Sha48]. As shown in multiple studies it is common to measure the time in bouts of activity [Byr16]. Based on the study of McVeigh et al., intervals for different intensity levels were defined from 0 to 5, 5 to 10, 10 to 20, 20 to 30, 30 to 60 and  $> 60$  minutes [McV16]. In addition, several intervals  $> 60$  minutes were defined to capture long periods of inactivity. The total time per day within these intervals were calculated. Furthermore, the mean of the barcode was also calculated. Due to the fact, that the amount of activity in different parts of the day can differ [vB18], most of the calculated features were also divided into different parts of the day. Altogether 92 features were computed for each barcode. A full list of features can be found in Appendix D.

### 3.3.2 Feature Extraction - Expert-Driven Approach

The results of the pure data-driven approach explained above may be difficult to describe with respect to a clinically relevant interpretation. Therefore the research group of the STAR Study [Gei17] defined a subset of features, that are most relevant from their perspective. In total, seven features were identified (Tab. 3.2). Four time-related features have been selected as well as three features related to variability and complexity of the barcode.

Feature Vector (Expert-Driven Approach)
Total Time Inactive (in % of Total Time)
Total Time Light Active (in % of Total Time)
Total Time High Active (in % of Total Time)
Longest Time Inactive (in % of Total Time)
Information Entropy
Information Entropy (Inactive vs. Active)
Information Entropy (High Active vs. Not High Active)

Table 3.2: List of features for the expert-driven approach

### 3.3.3 Dimensionality Reduction

Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality [Jol11]. This can be achieved by using the PCA, which transforms the  $d$ -dimensional feature set  $\mathbf{f}$  to a  $d'$ -dimensional feature set  $\mathbf{f}'$ , the principal components, while maximizing the variance of the transformed features (with  $d' \leq d$ , ideally  $d' \ll d$ ). To balance the variance of the features, the feature set  $\mathbf{f}$  was re-scaled before applying the PCA, such that all feature values were in the range  $[0, 1]$ .

The feature space  $\mathbf{f}'$  providing the maximized spread data is defined by the eigenvectors of the covariance matrix of the feature set  $\mathbf{f}$ . The eigenvectors can be calculated by solving the eigenvalue problem

$$\mathbf{S}\mathbf{u}_k = \lambda_k\mathbf{u}_k, \quad (3.2)$$

where

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \quad (3.3)$$

is the covariance matrix of the feature set  $\mathbf{f}$ .  $\mathbf{u}_k$  are the eigenvectors of  $\mathbf{S}$ ,  $\lambda_k$  are the eigenvalues of  $\mathbf{S}$ ,  $N$  is the amount of samples in  $\mathbf{f}$ ,  $\mathbf{x}_n$  are the feature values of the  $n^{\text{th}}$  sample and  $\bar{\mathbf{x}}$  are the mean vector of the features in  $\mathbf{f}$  [Bis06]. The eigenvectors and their corresponding eigenvalues were computed using a singular value decomposition (Eq. 3.4).

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (3.4)$$

The decomposition of  $\mathbf{S}$  results in  $\mathbf{\Sigma}$ , which has non-negative diagonal elements arranged in descending order of magnitude, and the orthogonal matrices  $\mathbf{U}$  and  $\mathbf{V}$ . The elements of  $\mathbf{\Sigma}$  are

called singular values of  $S$  and are the square roots of the eigenvalues of  $SS^T$ . The columns of  $U$  are eigenvectors of  $SS^T$  and the columns of  $V$  are eigenvectors of  $S^T S$ . By using

$$\mathbf{f}' = U^T \mathbf{f}, \quad (3.5)$$

$\mathbf{f}$  can now be transformed into  $\mathbf{f}'$  [Str93]. The resulting principal components in  $\mathbf{f}'$  are weighted linear combinations of the original features, uncorrelated and sorted in descending order such that the first component describes the most variance. The variance explained by every principal component can be calculated by the corresponding eigenvalue divided by the sum of all eigenvalues. The cut-off point for the dimensionality reduction was set to 90% of the cumulated explained variance. In the following the feature-set  $\mathbf{f}'$  is used for the cluster analysis. The PCA was implemented by using the scikit-learn python library [Ped11].

## 3.4 Clustering

Clustering defines a task of grouping similar data into the same cluster, such that each data point in a cluster is more similar to data points of the same cluster than those in other clusters [Bis06]. Grouping similar data points together can help to profile attributes of each group and can give insights into underlying patterns. There are different approaches to solve this task, that can basically be distinguished between hierarchical and partitional approaches [Jai99]. For the scope of this thesis, both approaches were used and compared.

### 3.4.1 Hierarchical Clustering

Hierarchical clustering can be divided into a divisive and an agglomerative approach. In the agglomerative approach each data point initially represents a cluster and then the clusters are successively merged until all clusters are merged. The divisive approach begins with all data points in a single cluster and will split a cluster in every iteration until a stopping criterion is met [Jai99]. For the scope of this thesis the agglomerative approach was used. In order to decide, which clusters should be merged a measure of dissimilarity between sets of data points is required. Therefore a distance metric and a linkage criterion which specifies the dissimilarity between sets as a function based on the distance metric is needed.

The euclidean distance was used as the distance metric. The distance  $d_{i,j}$  between the data points  $\mathbf{x}_i$  and  $\mathbf{x}_j$  was calculated with:

$$d_{i,j} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}. \quad (3.6)$$

As a linkage criterion the Ward's method was used [WJ63]. The method minimized the total within-cluster variance of the clusters being merged. The within-cluster variance  $\mathcal{W}$  of a cluster  $\mathcal{C}$  is computed by

$$\mathcal{W}(\mathcal{C}) = \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{c}_j\|^2, \quad (3.7)$$

where  $\mathbf{x}_i$  are the data points within the cluster and  $\mathbf{c}_j$  is the centroid of the cluster. The hierarchy of merged clusters is represented in a dendrogram, where the root is the unique cluster that gathers all the samples and the leaves are all clusters with only one data point in it. The dendrogram also contains distances between merged clusters. [Rok05]. Accordingly, the agglomerative clustering algorithm will not determine a certain number of clusters. Therefore, evaluation methods are needed to determine the optimal number of clusters. The agglomerative clustering algorithm was implemented by using the scikit-learn python library [Ped11].

### 3.4.2 Partitional Clustering

A partitional clustering algorithm produces one partition instead of a clustering structure, such as the dendrogram in the hierarchical approach. The partitional clustering produces cluster by optimizing an objective function. The most frequently used objective function is the squared error criterion, first suggested by MacQueen et al. [Mac67]. The squared error for the clustering  $\mathcal{L}$  of a data set  $\mathcal{X}$ , which contains  $K$  clusters, is

$$e^2(\mathcal{L}, \mathcal{X}) = \sum_{j=1}^K \sum_{i=1}^{n_j} \|\mathbf{x}_i^{(j)} - \mathbf{c}_j\|^2, \quad (3.8)$$

where  $\mathbf{x}_i^{(j)}$  is the  $i^{th}$  data point belonging to the  $j^{th}$  cluster and  $\mathbf{c}_j$  is the centroid of the  $j^{th}$  cluster. In the scope of this thesis the k-means clustering algorithm is used, which is based on a squared error criterion as the objective function. In contrast to the hierarchical approach, the number of clusters must be specified in advance. In a first step, the algorithm chooses  $k$  random initial cluster centers and assigns every data point to the closest cluster center. Then every cluster center will be

recalculated based on the current cluster memberships and every data point will be reassigned to the closest cluster center. The algorithm converges when there are no reassignments of data points to new cluster centers, which is equivalent to a minimal squared error [Jai99]. But based on the random initialization, the algorithm could lead to different clustering results, even to no result [Zha08]. Therefore, the algorithm was run ten times with different initializations and returns the best result regarding the objective function. For the evaluation, this algorithm was run with  $k \in [2, 20]$  initial cluster centers. The k-means algorithm was implemented by using the scikit-learn python library [Ped11].

## 3.5 Clustering Evaluation

A clustering evaluation can be based on internal and external criteria. Internal criteria for the quality of the clustering are typically based on an objective function with the goal of high intra-cluster and low inter-cluster similarity or distance, while the external criteria are based on previous knowledge about the data [Ren11].

### 3.5.1 Silhouette Method

As an internal criterion the silhouette method was used to interpret and validate the consistency within clusters [Rou87] and provides a graphical representation of how well each data point has been clustered. The silhouette value for each data point can be calculated with

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.9)$$

where  $a(i)$  is defined as the average distance from  $i$  to all points of the same cluster and  $b(i)$  is defined as the smallest average distance of  $i$  to all points of any other cluster of which  $i$  is not assigned to. The silhouette ranges from -1 to +1, where a high positive value indicates that a data point is well matched to its own cluster. As a distance measurement the euclidean distance (Eq. 3.6) was used.

The silhouette score  $s_C$  is a quality measurement of the clustering result and is defined as mean of all  $n_C$  silhouettes of a cluster  $C$ :

$$s_C = \frac{1}{n_C} \sum_{i \in C} s(i). \quad (3.10)$$

The average silhouette score of a clustering result was used for the evaluation of different initial numbers of clusters  $k$ .

### 3.5.2 External Evaluation

While the silhouette method is using the clustering result itself for the evaluation, there is also the possibility to evaluate the clustering result based on data that were not used for the clustering. Therefore, the clinical information of the patients was used as an external criterion to evaluate the quality of the clusters we identified based on the average silhouette score. For all patients in the data set, information was available on their FEV<sub>1</sub>, 6MWT and the outcome of the CAT<sup>TM</sup> (Fig. 3.1).

To evaluate the outcome of the clustering algorithm, the distribution of the clinical information within the clusters was calculated and an analysis of variance (ANOVA) with a significance level of 0.05 was applied, in order to determine whether the clusters differ significantly from each other [Edw93]. To describe the magnitude of the result, the effect size  $\eta^2$  was calculated by [Fer09]:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}, \quad (3.11)$$

where  $SS_{effect}$  is the sum of squares for the effect of interest and  $SS_{total}$  is the total sum of squares for all effects in the ANOVA. The sum of squares was calculated with Equation 3.12, where  $\bar{y}$  is the mean of the data points of interest  $y_i$ .

$$SS = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3.12)$$

The testing hypothesis of the ANOVA was:

$\mathcal{H}_0$ : No Difference between cluster means exists (e.g.  $\bar{x}_1 = \bar{x}_2 = \bar{x}_3$ )

$\mathcal{H}_1$ : Differences between cluster means exist (e.g.  $\bar{x}_1 \neq \bar{x}_2$  or  $\bar{x}_1 \neq \bar{x}_3$  or  $\bar{x}_2 \neq \bar{x}_3$ )

In cases of a significant difference in the cluster means, the Bonferroni post-hoc test was used to determine which clusters differ significantly. Therefore, the significance level of 0.05 needed to be divided by number of planned comparisons [Bla95]. The alpha level of the post-hoc test then needs to be less than the calculated quotient.



# Chapter 4

## Results

In the following sections the results of the cluster analysis will be presented. The results are divided into the data-driven and the expert-driven approach. For both approaches the result of the PCA will be shown as well as the results of the different clustering algorithms and the corresponding evaluation.

### 4.1 Data-Driven Approach

In the data-driven approach, 92 features were defined to describe the data-set (Appendix D). The following results are divided accordingly to the clustering algorithms.

#### PCA

The PCA identified 42 components, which accounted for 90.0% of the total variance of the data. The first three principal components accounted for 29.9% of the total variance (first component, 15.6%; second component, 8.2%; third component, 6.1%). The most relevant features for the first three principal components were: first component, the total time inactive; second component, the entropy between inactive and active in negative direction of the axis; third component, the time spent in < 60 minute bouts of inactivity in the third period of the day. The visualization of the first three principal components is shown in Figure 4.1.

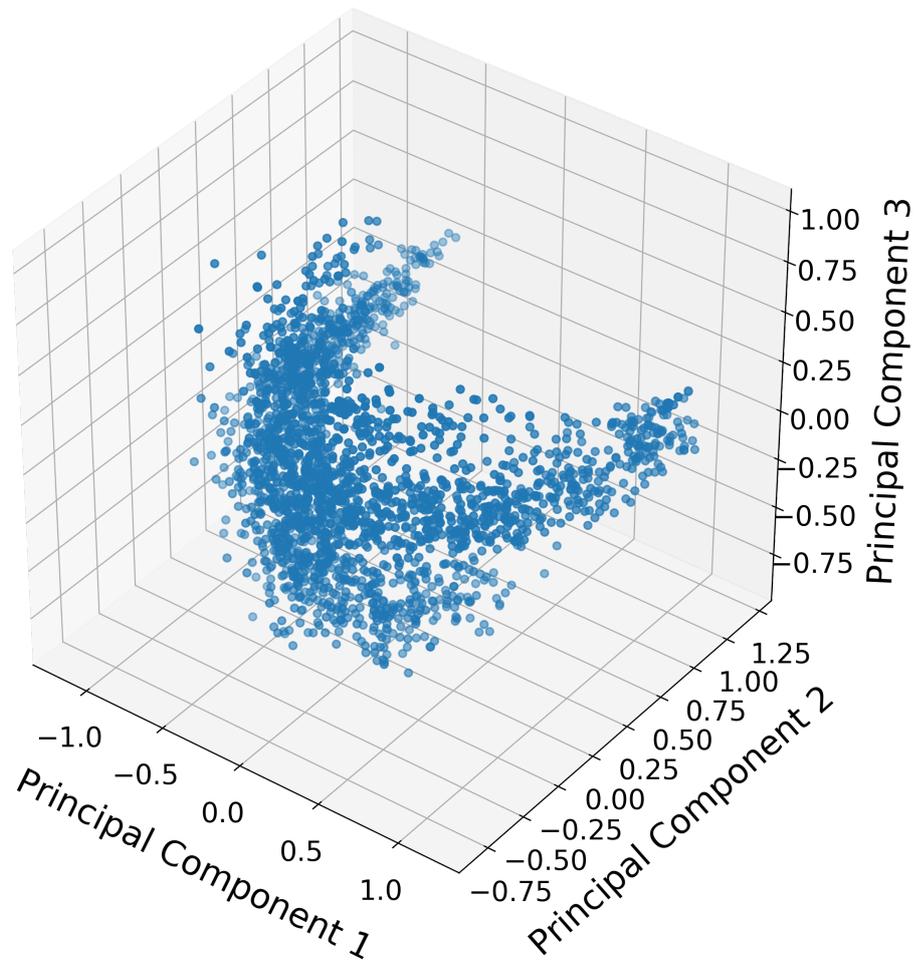


Figure 4.1: The result of the principal component analysis for the data driven approach. The first three principal components displayed here accounted for 29.9% of the total variance.

### 4.1.1 Hierarchical Clustering

The agglomerative clustering algorithm was evaluated for  $k$  clusters (with  $k \in [2, 20]$ ). For the evaluation, the dendrogram and the average silhouette score was used as well as the external evaluation. As shown in Figure 4.2,  $k = 2$  created the best clustering result with an average silhouette score of 0.101. The dendrogram did not show a clear differentiation between the clusters (Fig. 4.3). In Figure 4.4 the silhouette value of each data point is shown next to the result of the agglomerative clustering algorithm for  $k = 2$ . The clustering result with the first three principal components can be seen in Figure 4.5. The average FEV<sub>1</sub>, CAT<sup>TM</sup> score and 6MWT distance of the two clusters as well as the time in different activity intensities and the overall entropy of the barcode are shown in Table 4.1. The ANOVA showed a significant difference between the clusters for the FEV<sub>1</sub>, for the 6MWT and for the CAT<sup>TM</sup>. The following post-hoc tests revealed also a significant difference between the clusters for the FEV<sub>1</sub>, for the 6MWT and for the CAT<sup>TM</sup> (Tab. 4.1). Representative barcodes for the two clusters are shown in Figure 4.6 and Figure 4.7.

	Cluster 1	Cluster 2	P-value	Effect size
N	1420	836		
Time Inactive (in % of 10h)	61.7	26.6	-	-
Time Light Active (in % of 10h)	36.5	67.5	-	-
Time High Active (in% of 10h)	1.8	5.9	-	-
Entropy	0.89	0.95	-	-
FEV <sub>1</sub> (in %) **	51.1	58.1	<0.001	0.035
6MWT (in m) **	438.1	476.1	<0.001	0.033
CAT <sup>TM</sup> **	24	22	< 0.001	0.008

Table 4.1: Average characteristics of the different clusters of the agglomerative clustering algorithm within the data-driven approach

\*\* : Significant difference between Cluster 1 and Cluster 2 (p<0.001)

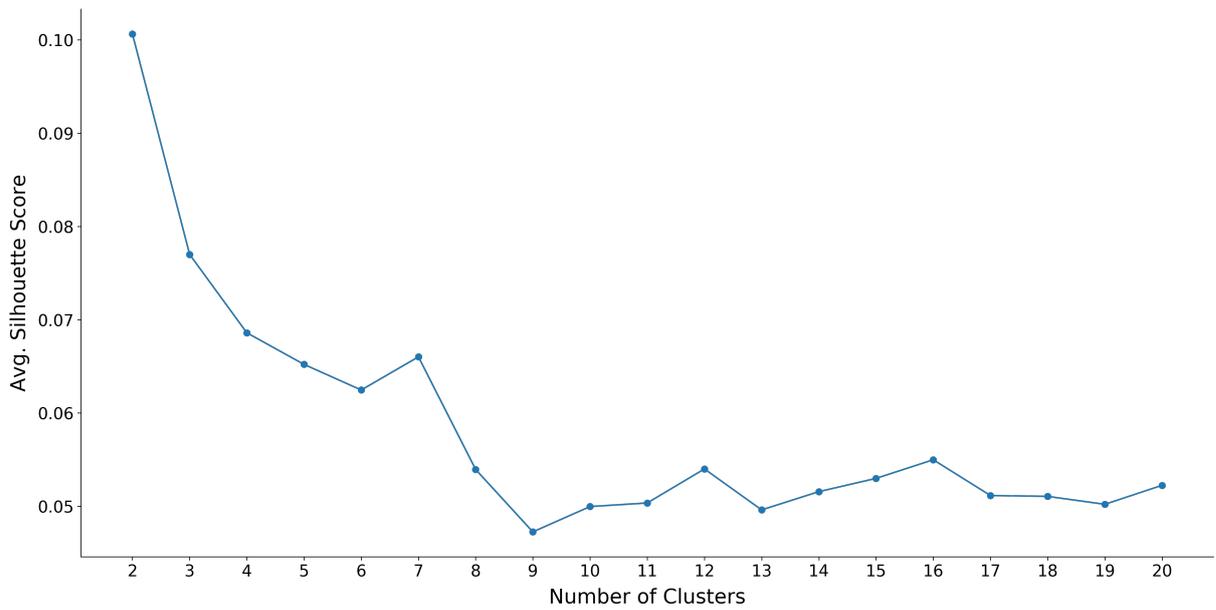


Figure 4.2: The average silhouette scores for a different number of clusters  $k$  ( $k \in [2, 20]$ ) of the agglomerative clustering within the data-driven approach

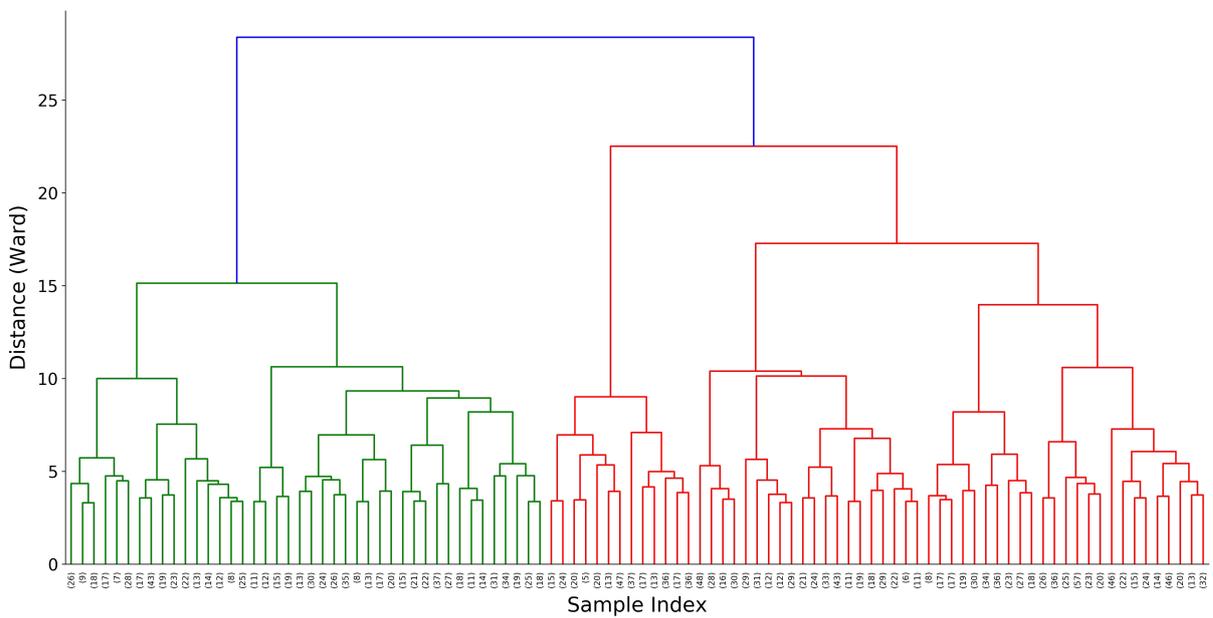


Figure 4.3: Dendrogram of the data-driven approach (only the last 100 merged cluster are shown)

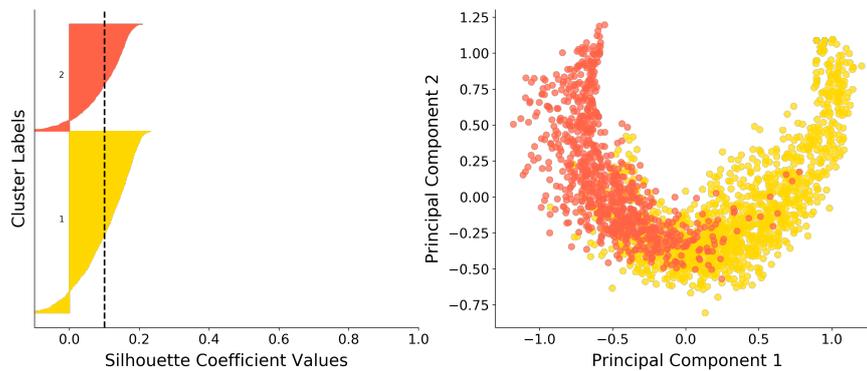


Figure 4.4: Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the agglomerative clustering within the data-driven approach on right side ( $k = 2$ )

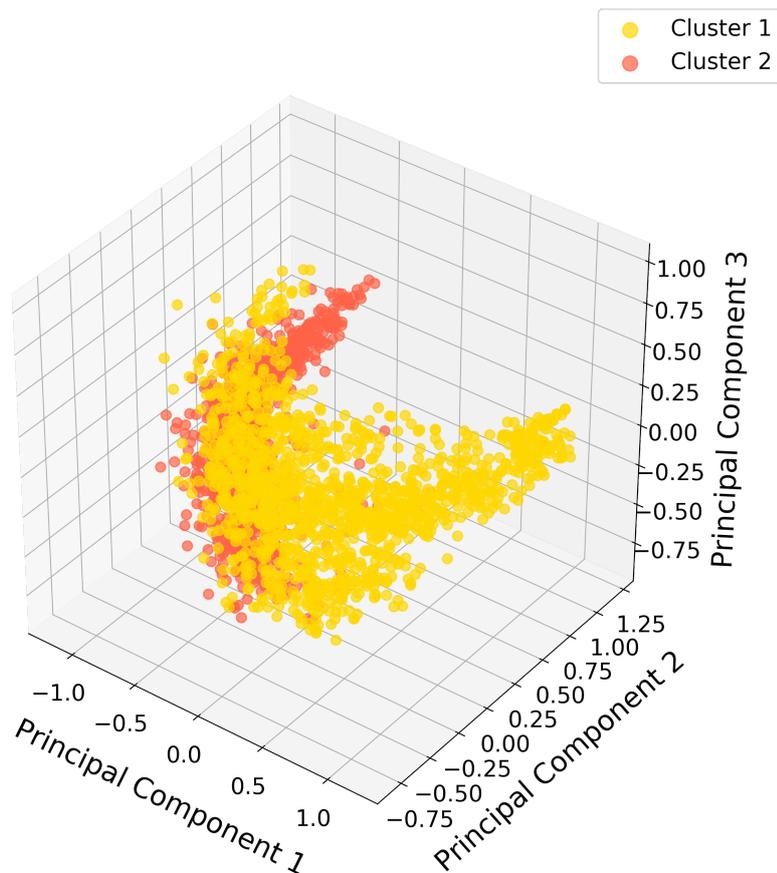


Figure 4.5: The agglomerative clustering result for the data-driven approach ( $k = 2$ )

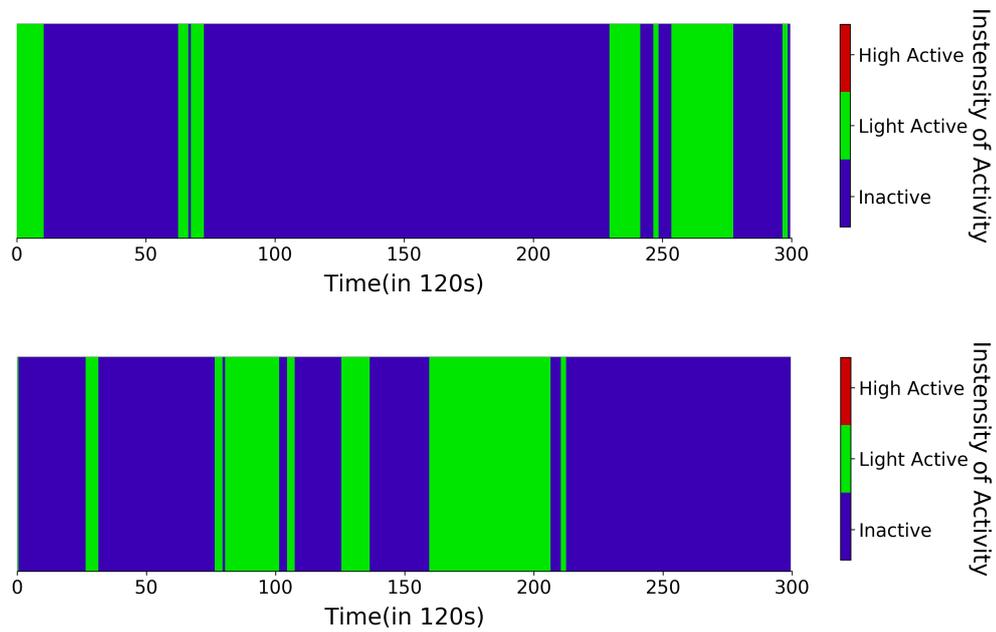


Figure 4.6: Representative barcodes for Cluster 1 of the agglomerative clustering result within the data-driven approach

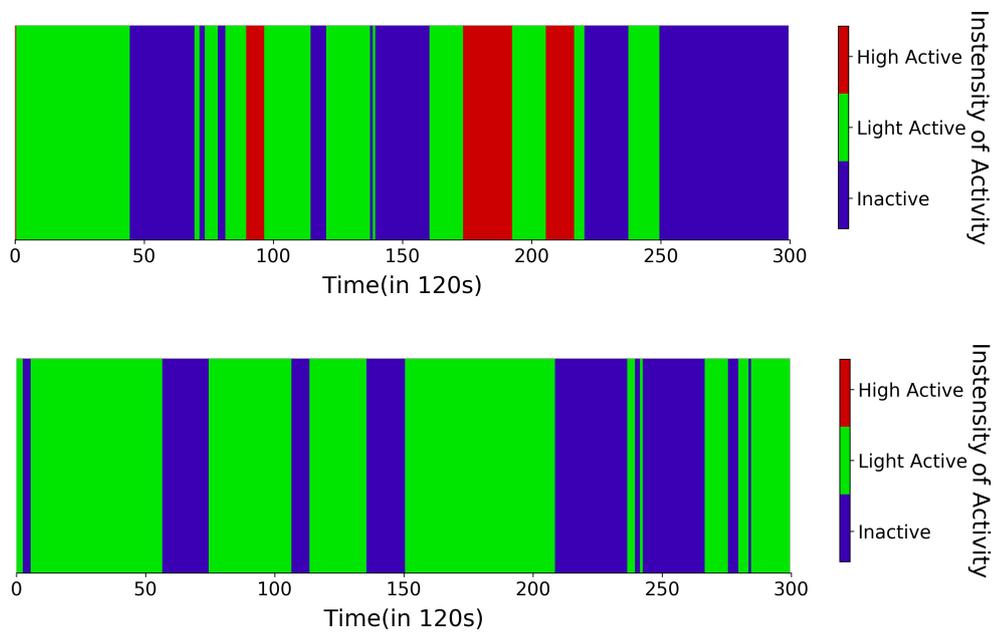


Figure 4.7: Representative barcodes for Cluster 2 of the agglomerative clustering result within the data-driven approach

### 4.1.2 k-Means

The k-means clustering algorithm was evaluated for  $k$  clusters (with  $k \in [2, 20]$ ). For the evaluation, the average silhouette score was used as well as the external evaluation. As shown in Figure 4.8,  $k = 2$  created the best clustering result with an average silhouette score of 0.121. In Figure 4.9 the silhouette value of each data point is shown next to the result of the k-means algorithm for  $k = 2$ . The clustering result with all three principal components can be seen in Figure 4.10. The average FEV<sub>1</sub>, CAT<sup>TM</sup> score and 6MWT distance of the four clusters as well as the time in different activity intensities is shown in Table 4.2. The ANOVA showed a significant difference between the clusters for the FEV<sub>1</sub>, for the 6MWT and for the CAT<sup>TM</sup>. The following post-hoc tests revealed also a significant difference between the clusters for the FEV<sub>1</sub>, for the 6MWT and for the CAT<sup>TM</sup> (Tab. 4.2). Representative barcodes for the two clusters are shown in Figure 4.11 and Figure 4.12.

	Cluster 1	Cluster 2	P-value	Effect size
N	1076	1180		
Time Inactive (in % of 10h)	70.57	28.71	-	-
Time Light Active (in % of 10h)	28.33	65.93	-	-
Time High Active (in% of 10h)	1.10	5.36	-	-
Entropy	0.83	0.98	-	-
FEV <sub>1</sub> (in %) **	49.2	57.8	<0.001	0.056
6MWT (in m) **	429.4	472.6	<0.001	0.044
CAT <sup>TM</sup> **	24	22	< 0.001	0.014

Table 4.2: Average characteristics of the different clusters of the k-mean algorithm within the data-driven approach

\*\* : Significant difference between Cluster 1 and Cluster 2 (p<0.001)

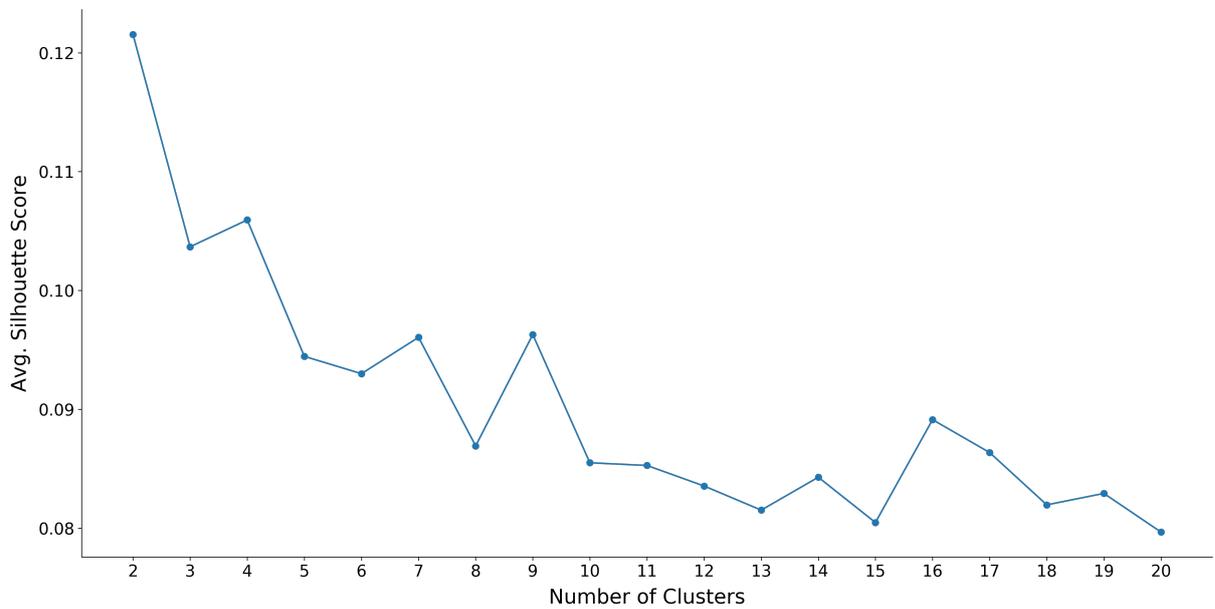


Figure 4.8: The average silhouette scores for a different number of clusters  $k$  ( $k \in [2, 20]$ ) of the agglomerative clustering within the data-driven approach

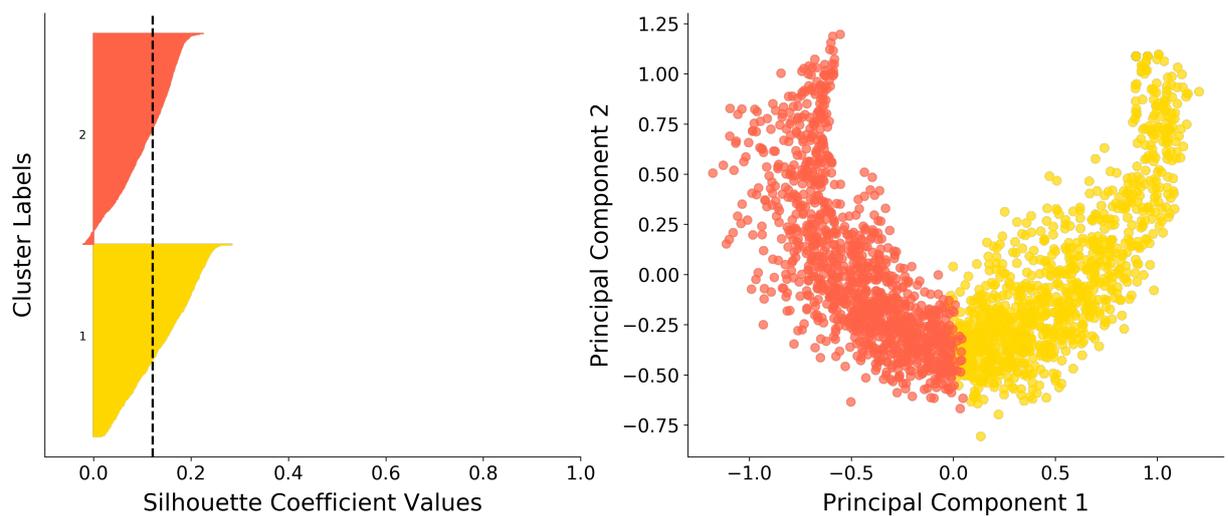


Figure 4.9: Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the k-means within the data-driven approach on the right side ( $k = 2$ )

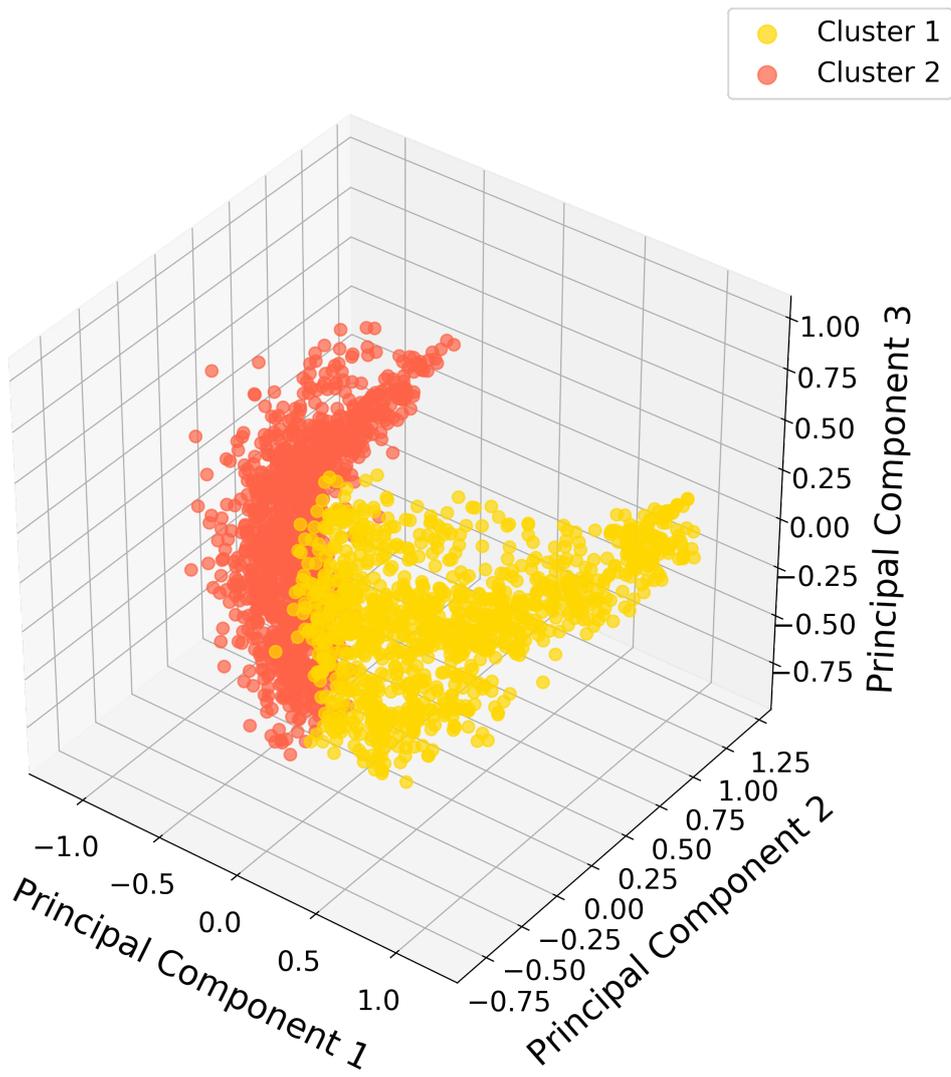


Figure 4.10: The k-means clustering result for the data-driven approach ( $k = 2$ )



## 4.2 Expert-Driven Approach

In the expert-driven approach, seven features were defined to describe the data-set. The following results are divided accordingly to the clustering algorithms.

### PCA

The PCA identified three principal components, which accounted for 94.8% of the total variance of the data (first component, 49.9%; second component, 25.7%; third component, 19.2%). The most relevant features of each component were: first component, the total time inactive; second component, the entropy between sedentary and non-sedentary behaviour in negative direction of axis; third component, the entropy between high and not high active behaviour 4.3. The visualization of the first three principal components is shown in Figure 4.13.

Feature	PC 1	PC 2	PC 3
Total Time Inactive (in % of Total Time)	0.65	-0.14	0.11
Total Time Light Active (in % of Total Time)	-0.56	0.18	-0.34
Total Time High Active (in % of Total Time)	-0.12	-0.06	0.34
Longest Time Inactive (in % of Total Time)	0.32	-0.06	0.10
Information Entropy	-0.21	-0.56	0.14
Information Entropy (Inactive vs. Active)	-0.03	-0.77	-0.41
Information Entropy (High Active vs. Not High Active)	-0.32	-0.18	0.74

Table 4.3: Correlation between the original features of the expert-driven approach and the resulting principal components (PC)

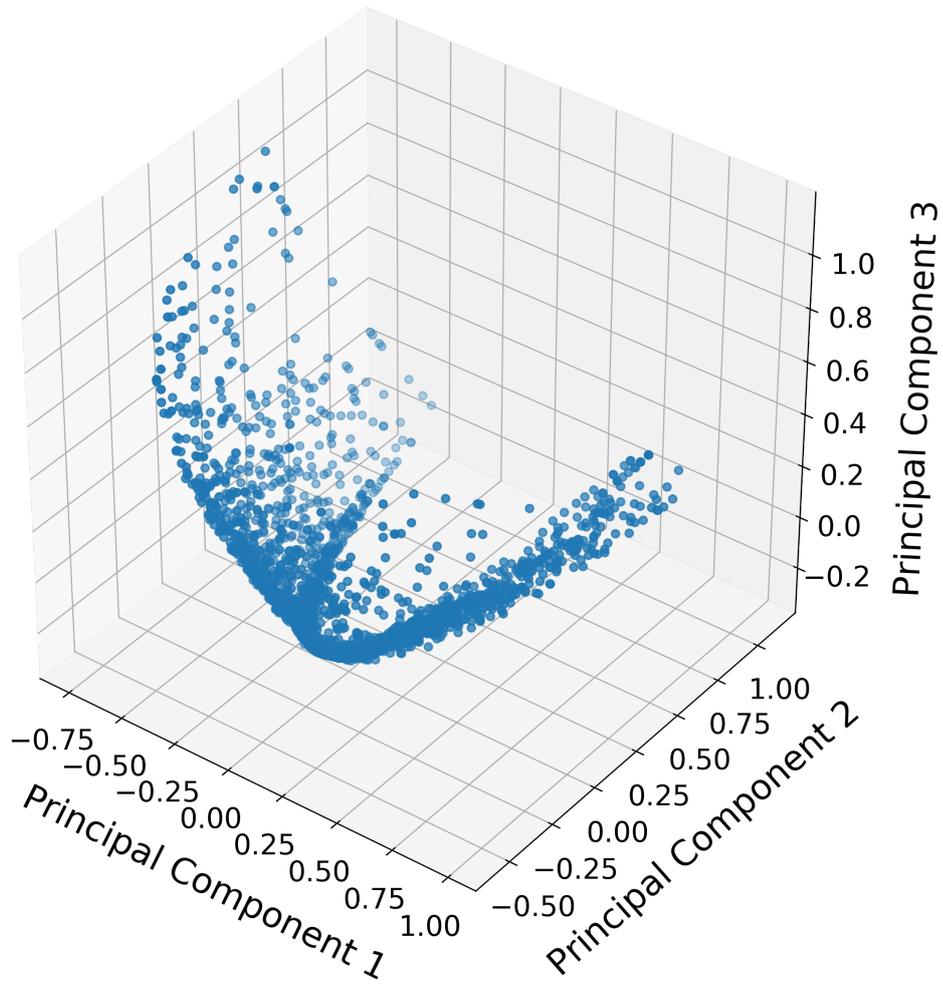


Figure 4.13: The result of the principal component analysis for the data driven approach. The first three principal components accounted for 94.8% of the total variance.

### 4.2.1 Hierarchical Clustering

The agglomerative clustering algorithm was evaluated for  $k$  clusters (with  $k \in [2, 20]$ ). For the evaluation, the dendrogram, the average silhouette score was used as well as the external evaluation. As shown in Figure 4.14,  $k = 2$  created the best clustering result with an average silhouette score of 0.405. This can also be supported by the dendrogram (Fig. 4.15). In Figure 4.16 the silhouette value of each data point is shown next to the result of the agglomerative clustering algorithm for  $k = 2$ . The clustering result with all three principal components can be seen in Figure 4.17. The average FEV<sub>1</sub>, CAT<sup>TM</sup> score and 6MWT distance of the four clusters as well as the time in different activity intensities and the overall entropy of the barcode are shown in Table 4.4. The ANOVA showed a significant difference between the clusters for the FEV<sub>1</sub>, for the 6MWT and for the CAT<sup>TM</sup>. The following post-hoc tests revealed also a significant difference between the two clusters for the FEV<sub>1</sub>, for the 6MWT and for the CAT<sup>TM</sup> (Tab. 4.4). Representative barcodes for every cluster are shown in Figure 4.18 - 4.19.

	Cluster 1	Cluster 2	P-value	Effect size
N	1973	282		
Time Inactive (in % of 10h)	42.9	89.2	-	-
Time Light Active (in % of 10h)	53.3	10.7	-	-
Time High Active (in% of 10h)	3.8	0.1	-	-
Entropy	0.98	0.47	-	-
FEV <sub>1</sub> (in %) **	54.9	45.3	<0.001	0.03
6MWT (in m) **	457.7	412.53	<0.001	0.02
CAT <sup>TM</sup> **	23	26	< 0.001	0.02

Table 4.4: Average characteristics of the different clusters of the agglomerative algorithm within the expert-driven approach

\*\* : Significant difference between Cluster 1 and Cluster 2 (p<0.001)

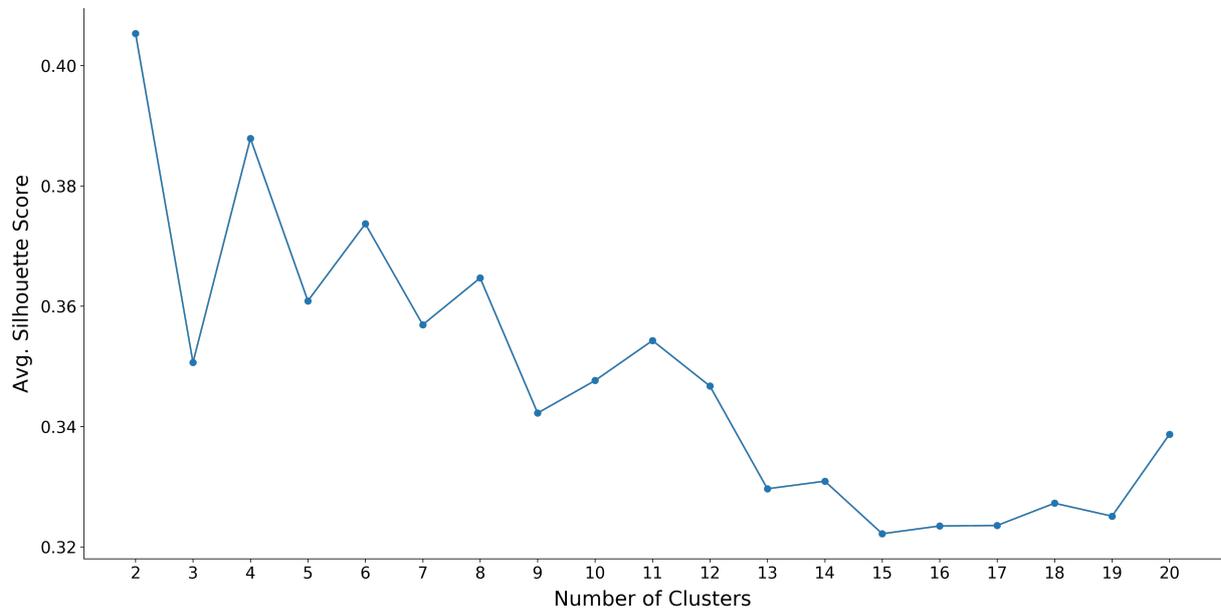


Figure 4.14: The average silhouette scores for a different number of clusters  $k$  ( $k \in [2, 20]$ ) of the agglomerative clustering within the expert-driven approach

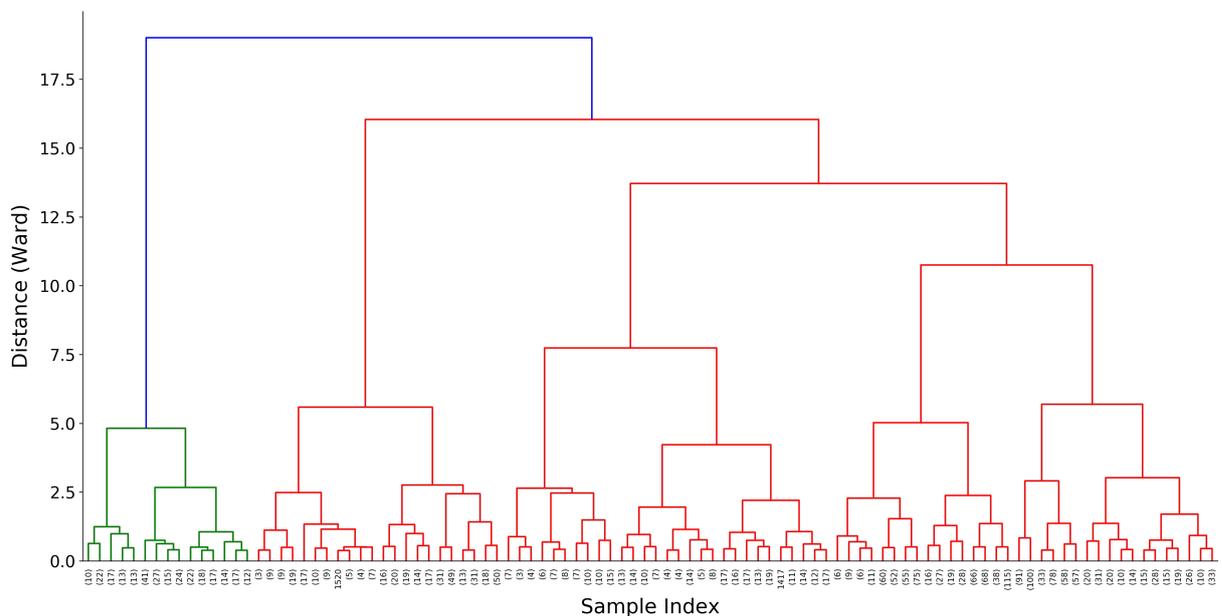


Figure 4.15: Dendrogram of the expert-driven approach  
(only the last 100 merged cluster are shown)

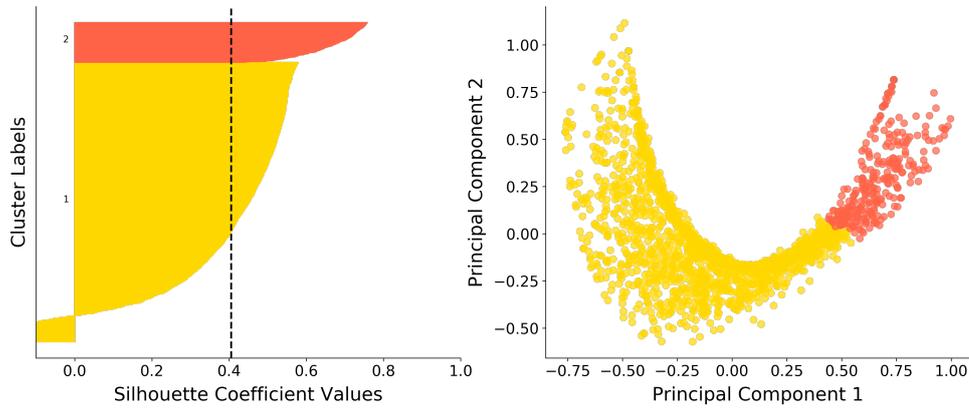


Figure 4.16: Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the agglomerative clustering within the expert-driven approach on the right side ( $k = 2$ )

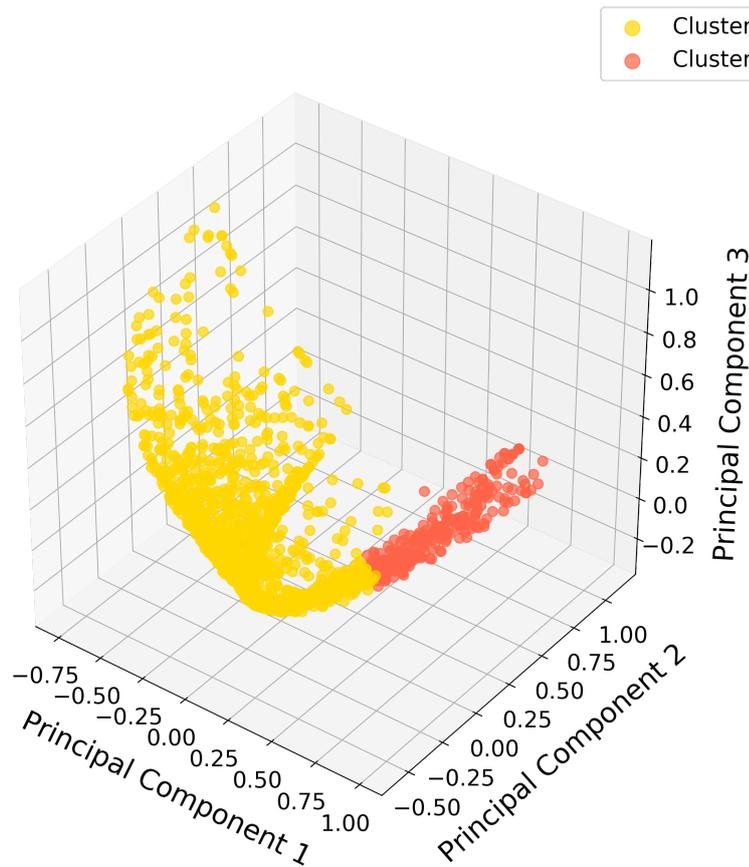


Figure 4.17: The agglomerative clustering result for the expert-driven approach ( $k = 2$ )

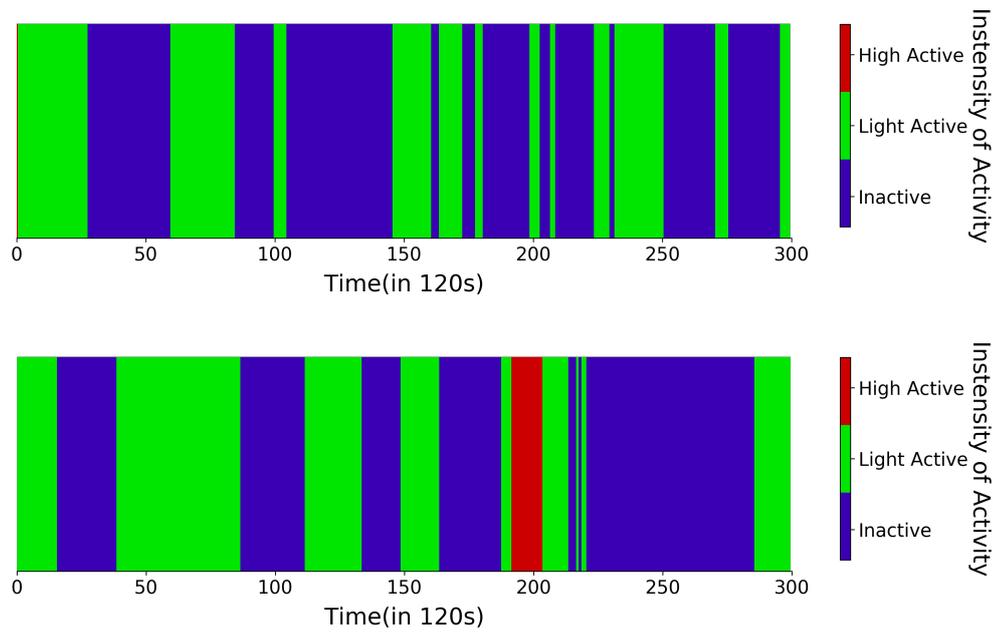


Figure 4.18: Representative barcodes for Cluster 1 of the agglomerative clustering result within the expert-driven approach

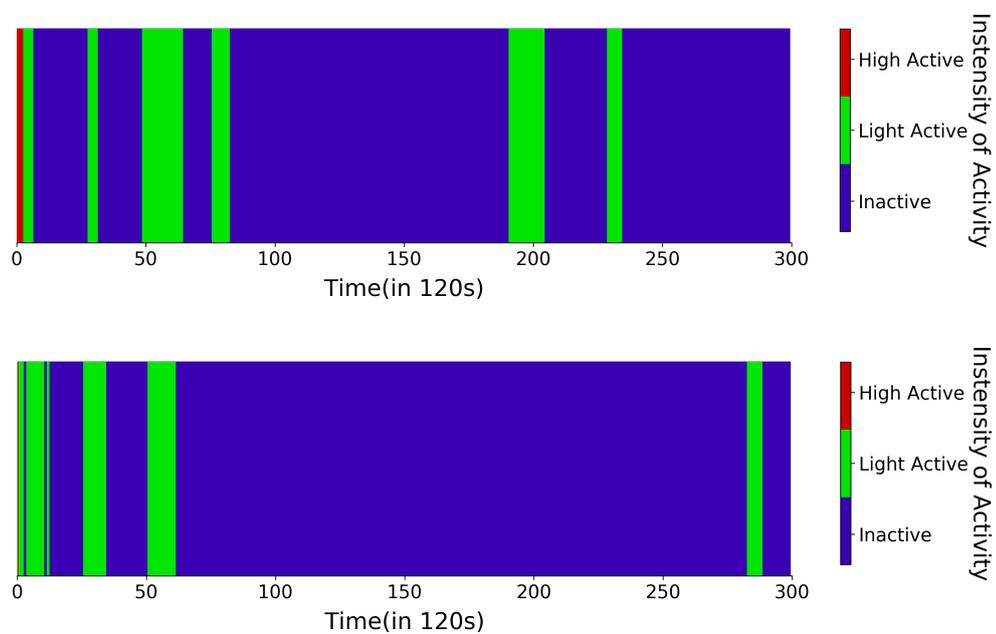


Figure 4.19: Representative barcodes for Cluster 2 of the agglomerative clustering result within the expert-driven approach

### 4.2.2 k-Means

The k-means clustering algorithm was evaluated for  $k$  clusters (with  $k \in [2, 20]$ ). For the evaluation, the average silhouette score was used as well as the external evaluation. As shown in Figure 4.20,  $k = 4$  created the best clustering result with an average silhouette score of 0.435. In Figure 4.21 the silhouette value of each data point is shown next to the result of the k-means algorithm for  $k = 4$ . The clustering result with all three principal components can be seen in Figure 4.22. The average FEV<sub>1</sub>, CAT<sup>TM</sup> score and 6MWT distance of the four clusters as well as the time in different activity intensities is shown in Table 4.5. The ANOVA showed a significant difference between the clusters for the FEV<sub>1</sub>, for the 6MWT and for the CAT<sup>TM</sup>. The following post-hoc tests revealed also a significant difference between all clusters for the FEV<sub>1</sub> and for the 6MWT. For the CAT<sup>TM</sup> there were significant differences between all clusters except between the clusters 2 and 4 (Tab. 4.5). Representative barcodes for every cluster are shown in Figure 4.23 - 4.26.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	P-value	Effect size
N	1103	420	423	309		
Time Inactive (in % of 10h)	52.6	16.0	85.3	28.6	-	-
Time Light Active (in % of 10h)	46.2	81.8	14.3	54.9	-	-
Time High Active (in% of 10h)	1.2	2.2	0.4	16.5	-	-
Entropy	1.01	0.72	0.58	1.29	-	-
FEV <sub>1</sub> (in %) **	52.8	57.2	46.1	62.3	<0.001	0.07
6MWT (in m) **	443.3	476.5	416.3	498.3	<0.001	0.07
CAT <sup>TM</sup> *	23	22	25	21	< 0.001	0.03

Table 4.5: Average characteristics of the different clusters of the agglomerative algorithm within the expert-driven approach

\*\* : Significant difference between all clusters ( $p < 0.001$ )

\* : Significant difference between all clusters except between Cluster 2 and Cluster 4 ( $p < 0.001$ )

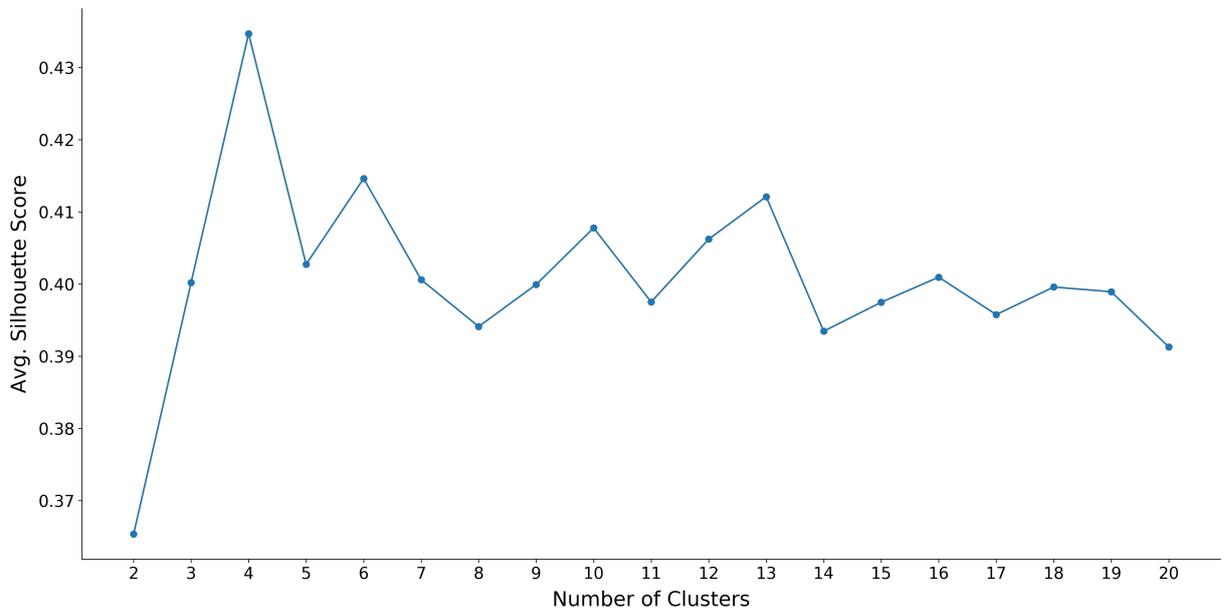


Figure 4.20: The average silhouette scores for a different number of clusters  $k$  ( $k \in [2, 20]$ ) of the k-means clustering within the expert-driven approach

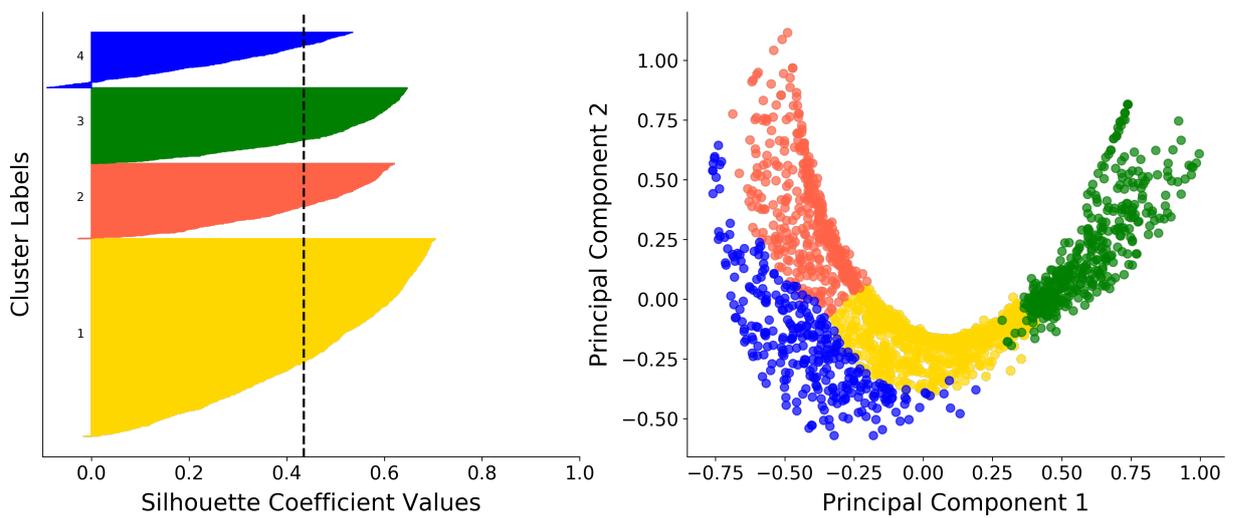


Figure 4.21: Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the agglomerative clustering within the expert-driven approach on the right side ( $k = 4$ )

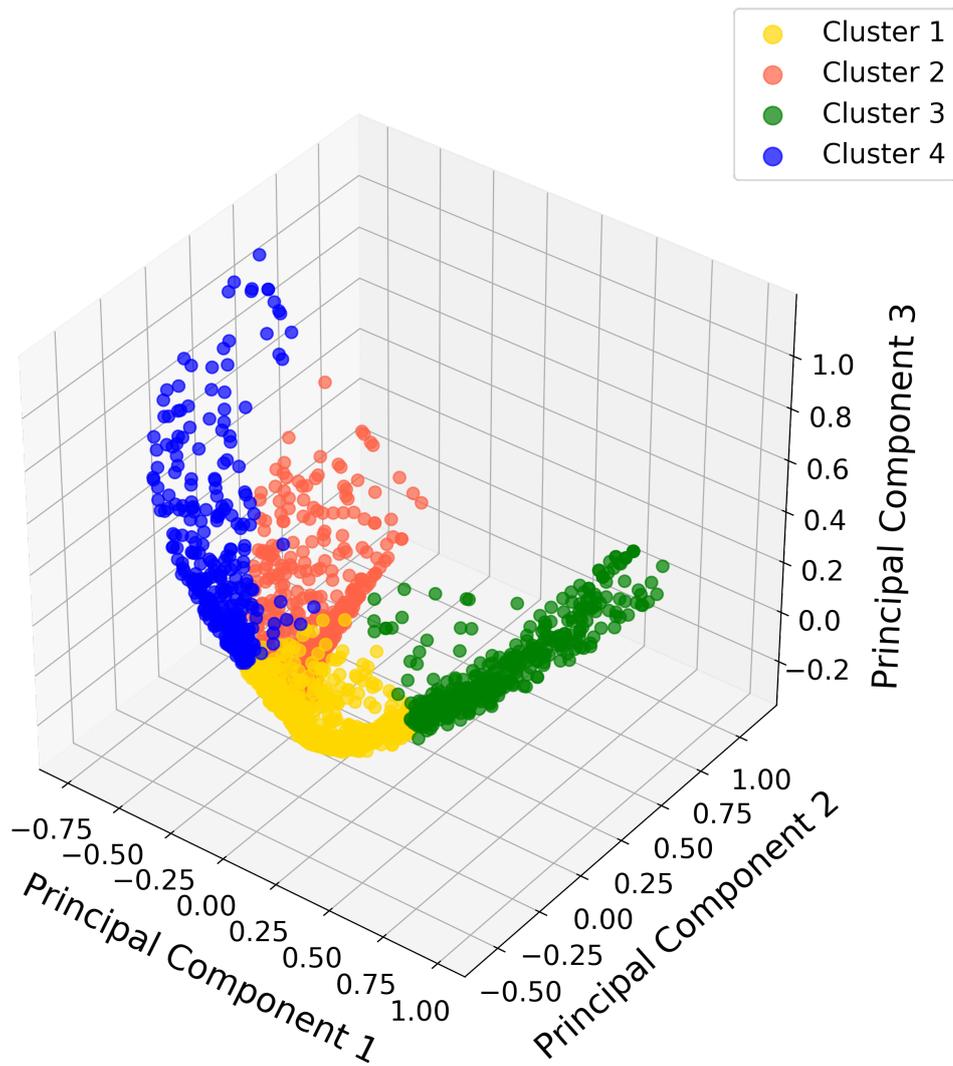


Figure 4.22: The k-means clustering result for the expert-driven approach ( $k = 4$ )



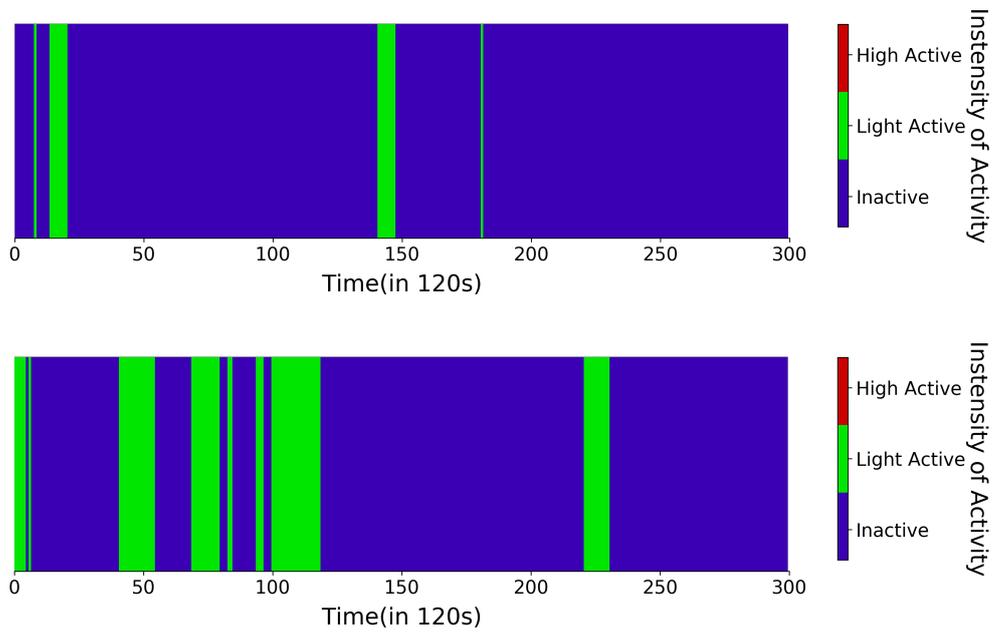


Figure 4.25: Representative barcodes for Cluster 3 of the k-means clustering result within the expert-driven approach

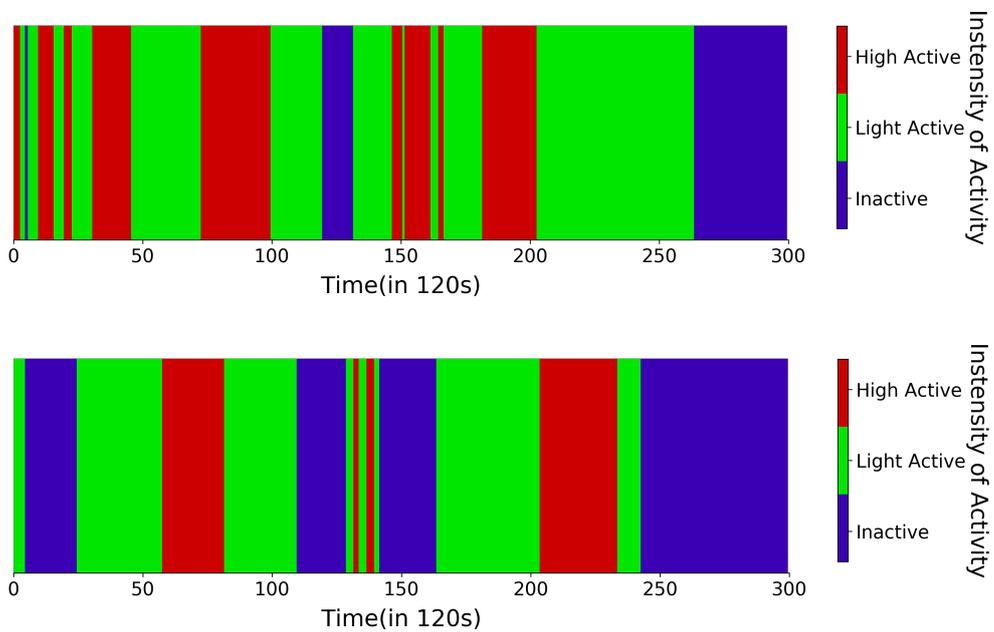


Figure 4.26: Representative barcodes for Cluster 4 of the k-means clustering result within the expert-driven approach



# Chapter 5

## Discussion

Analyzing the PA of COPD patients in more detail can provide additional insights into the effects of increased PA and the effectiveness of rehabilitation strategies. Based on the PA data provided by the research group of the STAR Study, a barcode representation of the data was computed. For the feature extraction two different approaches were used to extract features. To identify clusters, a hierarchical and a partitional clustering algorithm were applied. The results will be discussed in the following section.

### 5.1 Data Pre-Processing

Based on the barcode concept introduced by Paraschiv-Ionescu et al. [PI12], the activity counts measured by the Actigraph were transformed into a barcode with three different PA intensities and a time resolution of two minutes. The features that were extracted from the barcode could be clustered into subgroups with significant differences in the FEV<sub>1</sub>, the 6MWT distance and the CAT<sup>TM</sup> score (Tab. 4.1, 4.2, 4.4, 4.5). These differences could also be observed in the representative barcodes for all clustering results shown in the previous section. Therefore, the barcode concept can be considered as a reliable approach to represent PA throughout a day. Furthermore, a moving median filter with the window size of 15 data points was used to smooth the barcode and to remove artifacts and scattered peaks in activity. Even though a smoothed barcode could simplify the cluster analysis, the filter also removes information that could provide additional insights. Future investigations will need to evaluate different time resolutions, definitions of PA states and filters to validate or challenge the results of this work.

## 5.2 Data-Driven vs. Expert-Driven Approach

To identify clusters within a data set, features that describe the variance and properties of the original data need to be extracted. In machine learning, this process is called feature extraction. It is a dimensionality reduction process, where the high dimensional raw data is reduced to a lower dimensional feature vector that can still accurately describe the original data set. In the scope of this thesis, two approaches for the feature extraction were used.

In the data-driven approach, a variety of features were extracted. Features were computed, that are used in several studies [Byr16] such as time in bouts of activity with different intensity as well as features that may be able to describe the severity of the disease in different parts of the day [vB18]. This resulted in a total of 92 extracted features (Appendix D). A PCA compresses the feature vector to a lower dimension by capturing directions of maximal variance. When the variance distribution is low (high) within in the feature vector, more (less) principal components are needed to describe the variance of the original data. The PCA identified 42 principal components, to describe  $> 90\%$  of the total variance of the data. Furthermore, the principal components are linear combinations of all original features, which makes the result difficult to interpret, especially with high dimensional feature vectors. Because of the high dimensionality, it is not trivial to perform a cluster analysis on the 42 principal components. Kriegel et al. described four main problem statements for the clustering of high dimensional data [Kri09].

Firstly, high dimensions are hard to handle for the human intuition, impossible to visualize, and due to the exponential growth of possibilities with each dimension, a complete enumeration of all subspaces becomes increasingly difficult or even impossible. With 92 total features and 42 identified principal components the first described problem might affect the cluster analysis and also complicate the interpretation of the results.

Secondly, the concepts distance and similarity become less meaningful as the number of dimensions grows, since the distance between any two points in a given data-set converges to 0 for increasing dimensionality. This might have had a particularly great impact on the clustering result, since both clustering algorithms used the euclidean distance.

Thirdly, given a large number of features some features will usually not be meaningful for the assigned cluster. When clusters are only defined by a few of the available features, many features may interfere with the efforts to find these clusters. Irrelevant features can also be related to as noise. If there is not only global noise but given sets of features are noisy only with respect to certain sets of samples, different clusters only exist in different subsets of features. The described problem has not been observed in the cluster analysis for the data-driven approach, but this may also be a reason for the poor cluster separation.

Fourthly, given a large number of features, it is also likely that some features are correlated, which makes it more difficult to distinguish clusters. Especially in the data-driven approach, a variety of time-related features was extracted and it was not ruled out that some features might be correlated.

All this described problems are also often referred to as curse of dimensionality. To avoid the problems and to make the results more interpretable the research group of the STAR Study [Gei17] was asked to choose a subset of features with high relevance based on their expertise and experience. Eventually, the results should be interpretable in order to be useful for the assessment of PA in COPD patients.

In this expert-driven approach, seven features were identified to describe the data-set (Tab. 3.2). As described in the result section, the PCA identified three principal components, to describe  $> 90\%$  of the total variance of the data. With three principal components and seven features describing those principal components, the results are better interpretable and the problems of high-dimensional data described above do not apply.

### 5.3 Hierarchical vs. Partitional Clustering

Agglomerative clustering as a hierarchical clustering approach and k-means as a partitional clustering approach were separately applied to the data-driven and expert-driven feature vector. Since both algorithms are based on distance measurement, it is not surprising that the clustering results for the high-dimensional data-driven approach did not yield good results [Kri09]. Neither the dendrogram (Fig. 4.3), nor the silhouette scores (Fig. 4.2, Fig 4.8) supported a clear cluster separation. The possible reasons for this have been described in the previous section. Therefore, only the results of the expert-driven approach are discussed in more detail.

The agglomerative clustering algorithm computed a hierarchy of partitions based on Ward's criterion which is displayed in the dendrogram. (Fig. 4.3). Therefore, the dendrogram provides information on how the clusters were constructed and it visualizes the distances between individual clusters. But in order to decide for a number of clusters, an evaluation method like the silhouette score must be considered. As shown in Figure 4.14,  $k = 2$  created the best clustering result with an average silhouette score of 0.405.

The k-means clustering algorithm provides one final partition for a pre-defined number of clusters  $k$ . Therefore, several initial number of clusters were tested and evaluated by the silhouette method. As shown in Figure 4.20,  $k = 4$  created the best clustering result with an average silhouette score of 0.435.

While an average silhouette score close to +1.0 indicates a good clustering result [Rou87], the scores presented are close to +0.4. This can be explained by the fact that the data set does not form well-separable clusters, but there are still tendencies which can be detected by a cluster analysis (Fig. 4.17, Fig. 4.22). However, comparing the agglomerative clustering and k-means clustering based on the silhouette score, the k-means algorithm achieved a slightly better result with an average silhouette score of 0.435 and will be the foundation of the interpretation in the following section.

## 5.4 Interpretation of the Clustering Results

The goal of this thesis was to distinguish daily profiles of PA for COPD patients. Even though the evaluation based on the silhouette score was not in the range of a good clustering result [Rou87], an actual interpretation of the results could provide useful information from a clinical perspective. Based on the results of the expert-driven k-means algorithm, the average FEV<sub>1</sub> score, the CAT<sup>TM</sup> result and the 6MWT distance were calculated for all four clusters as well as entropy and the average time in different activity intensities (Tab. 4.5).

Cluster 1 ( $n = 1103$ ) has a balanced proportion of inactive (52.6%) and light active (46.2%) time during the day with a very small amount high-intensity activity (1.2%). Cluster 1 is therefore called *Sedentary Movers*.

Cluster 2 ( $n = 420$ ) spends most of their time in light-intensity PA (81.8%) and is therefore called *Restless Movers*.

Cluster 3 ( $n = 423$ ) is mostly inactive (85.3%) and spends less than 15% of the day in light to high-intensity PA. It is therefore called *Coach Potatoes*.

Cluster 4 ( $n = 309$ ) has a relatively high amount of high-intensity PA (16.5%) throughout the day and also spends 54.9% of the day in light intensities. Based on these characteristics cluster 4 is called *High Active Movers*.

As shown in Table 4.5 the differences in the average FEV<sub>1</sub> score and 6MWT distance are significant between all clusters, while the CAT<sup>TM</sup> showed significant differences between all clusters except between the *Restless Movers* and *High Active Movers*. Also supported by the effect size  $\eta^2$  of 0.07 for the FEV<sub>1</sub> score and 6MWT distance, a larger difference can be found in those parameters than in the CAT<sup>TM</sup> with an effect size  $\eta^2$  of 0.03. That could be attributed to

the fact that the FEV<sub>1</sub> and 6MWT are parameters measuring the physical capability, while the CAT<sup>TM</sup> score is based on a questionnaire and could be biased by the patient.

The highest average FEV<sub>1</sub> score and 6MWT distance as well as the lowest CAT<sup>TM</sup> score can be found in the *High Active Movers* cluster (Tab. 4.5). Compared to the other clusters, barcodes of the *High Active Movers* also have the highest entropy. The lowest average FEV<sub>1</sub> score and 6MWT distance as well as the lowest entropy value can be found in the *Couch Potatoes* cluster. This result not only underlines the importance of high-intensity PA, but also reveals a correlation between the entropy of a barcode and the severity of COPD. A high entropy score is correlated to a high number of changes in the intensity of activity [Sha48]. This result therefore supports the findings of Healy et al. [Hea11] that breaking up sedentary time can have a beneficial impact on the patients health.

Compared to the findings of Lee et al. [Lee13], the cluster analysis of the expert-driven approach did not only identified one 'active' and one 'inactive' cluster, but was able to further divide the 'active' and 'inactive' clusters. While the *Coach Potatoes* and *Sedentary Movers* could be described as 'inactive' because they spend most of their day inactive, the *Restless Movers* and *High Active Movers* could be described as 'active', because they spend most of their day in light to high-intensity PA. Mesquitea et al. also found a very active and a very inactive cluster as well as three clusters in between based daily physical activity measures [Mes17]. The most inactive cluster they found could also be characterized by worse airflow limitations and a higher amount of inactivity than other clusters, while the most active cluster could be characterized by a relatively high amount of high-intensity PA. Those findings underline the fact, that there are different PA behaviour in COPD patients, which may also need a individualized intervention strategy. Mostly inactive clusters like the *Coach Potatoes* and *Sedentary Movers* probably benefit from interventions focusing not only on increasing the time in moderate-to-vigorous PA, but also on reducing the time in sedentary behaviour. Potential approaches to increase PA in COPD are described in a recent systematic review by Mantoani et al. [Man16]. Especially the *Coach Potatoes* might find it difficult to increase the time in higher intensities, which suggest that it might be more realistic to increase the time in light-intensity PA [Spr15]. In contrast to this, *High Active Movers* would probably benefit more from a PR that focuses on maintaining their high PA level throughout the day. Nevertheless, further research is needed to identify distinct daily profiles of COPD patients and develop appropriate rehabilitation strategies.

## 5.5 Strengths and Limitations

With a total of 5083 individual day measurements of PA in COPD patients, the results in this thesis are based on a large sample size. However, the days had to be modified to improve comparability and therefore only the first ten hours were considered. In future studies, other approaches should be evaluated. The cluster analysis could identify clusters with significant differences in the FEV<sub>1</sub> score and 6MWT distance between all clusters and therefore indicates a relationship between different PA behaviour and severity of COPD. However, additional information regarding comorbidities were not available and therefore the influence of other diseases on the patients PA could not be ruled out. A further limitation of the results was the inability to find meaningful features, which could have lead to better separated the clusters. Even tough the expert-driven approach provided a interpretable result, it has not achieved a distinct separation of the clusters. Future Work may evaluate which features are most relevant for cluster analysis as well as for clinically relevant interpretations.

# Chapter 6

## Conclusion and Outlook

The purpose of this thesis was to analyze the provided data-set with a high time resolution to identify and distinguish daily activity profiles of COPD patients.

The presented literature provided a variety of features, that could describe a patient's day. However, there were no features with such a high variance, that a pure data-driven feature vector could be compressed to a significantly lower dimension. Therefore, the research group of the STAR Study defined a subset of features with high relevance based on their expertise and experience. The identified expert-features could be compressed by the PCA to a three dimensional feature space, that described nearly 95% of the total variance of the expert feature vector.

The k-means clustering algorithm provided the best clustering result considering the average silhouette score. The average clinical parameter showed a significant difference between the identified clusters. Furthermore, a correlation between entropy and the severity of COPD as well as a correlation between the time spent in high intensities and the severity of COPD could be observed. Especially the the entropy of the barcode could be an interesting feature for the clinical context and supports that breaking up sedentary time can have a beneficial impact on the patients health.

Future studies may use these findings to determine typical daily profiles of COPD patients and to investigate the effect of PR on them. Mostly inactive patients may benefit more from a rehabilitation program focusing on reducing sedentary time while increasing the amount of light-intensity PA during the day because it may be difficult for such inactive patients to increase the amount of high-intensity PA. Patients with a considerable high amount of high-intensity PA during the day may benefit more from a PR focusing on strategies that help the patients to maintain this high amount of activity. A better understanding of the daily PA behaviour of COPD patients is therefore essential to improve and to individualize rehabilitation strategies. The

presented approach of analyzing PA of COPD patients may also be valuable for other diseases like Parkinson's disease, where a detailed analysis of daily PA may reveal new insights on the effectiveness of the therapy.

# Appendix A

## Glossar

<b>PA</b>	Physical Activity
<b>COPD</b>	Chronic Obstructive Pulmonary Disease
<b>PR</b>	Pulmonary Rehabilitation
<b>PCA</b>	Principal Component Analysis
<b>PC</b>	Principal Component
<b>FEV<sub>1</sub></b>	Forced Expiratory Volume in one second
<b>FVC</b>	Forced Vital Capacity
<b>CAT<sup>TM</sup></b>	COPD Assessment Test
<b>6MWT</b>	6-Minute Walk Test
<b>ANOVA</b>	Analysis of Variance



# Appendix B

## Patents

### B.1 US8337431B2

<b>Title</b>	Collecting activity and sleep quality information via a medical device
<b>Publication Number</b>	US8337431B2
<b>Publication Date</b>	2012-12-25
<b>Inventor(s)</b>	Kenneth T. Heruth, Keith A. Miesel
<b>Current Assignee</b>	Medtronic Inc
<b>Abstract</b>	<p>A device, such as an implantable medical device (IMD) or a programming device, determines when a patient is attempting to sleep. When the device determines that the patient is attempting to sleep, the device determines values for one or more metrics that indicate the quality of a patient's sleep based on at least one physiological parameter of the patient. When the device determines that the patient is not attempting to sleep, the device periodically determines activity levels of the patient. Activity metric values may be determined based on the determined activity levels. A clinician may use sleep quality information and patient activity information presented by a programming device to, for example, evaluate the effectiveness of therapy delivered to the patient by the medical device.</p>

## B.2 US8543185B2

<b>Title</b>	Activity monitoring systems and methods of operating same
<b>Publication Number</b>	US8543185B2
<b>Publication Date</b>	2013-09-24
<b>Inventor(s)</b>	Shelten Gee Jao Yuen, James Park, Eric Nathan Friedman
<b>Current Assignee</b>	Fitbit Inc
<b>Abstract</b>	The present inventions, in one aspect, is an activity monitoring system comprising a fixture having size/shape adapted to couple to a location on the user's body and a particular signature; and a portable monitoring device adapted to detect the fixture's particular signature. The monitoring device includes a housing that is adapted to engage the fixture; an activity sensor, disposed in the housing, to detect activity of the user and to generate data which is representative of the activity of the user; and processing circuitry, disposed in the housing, to calculate an activity-related quantity of the user, wherein the processing circuitry determines the monitoring device is engaging the fixture by detecting the fixture's particular signature and calculates the activity-related quantity.

## B.3 US9737261B2

<b>Title</b>	Wearable athletic activity monitoring systems
<b>Publication Number</b>	US9737261B2
<b>Publication Date</b>	2017-08-22
<b>Inventor(s)</b>	Aurel Coza, Christian Dibenedetto, Ian Michael Munson
<b>Current Assignee</b>	adidas AG
<b>Abstract</b>	A sensor garment for monitoring an individual engaged in an athletic activity includes a garment formed of textile material, and a sensor module inseparably coupled to the textile material of the garment. The sensor module includes a single-purpose sensor configured to sense a single characteristic, and a radio antenna configured to transmit data generated by the single-purpose sensor. The sensor module includes no external port.

# **Appendix C**

## **COPD Assessment Test**

Your name:

Today's date:



## How is your COPD? Take the COPD Assessment Test™ (CAT)

This questionnaire will help you and your healthcare professional measure the impact COPD (Chronic Obstructive Pulmonary Disease) is having on your wellbeing and daily life. Your answers, and test score, can be used by you and your healthcare professional to help improve the management of your COPD and get the greatest benefit from treatment.

For each item below, place a mark (X) in the box that best describes you currently. Be sure to only select one response for each question.

**Example:** I am very happy (0)  (1) (2) (3) (4) (5) I am very sad

	SCORE
<div style="display: flex; justify-content: space-between;"> <span><b>I never cough</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>I cough all the time</b></span> </div>	↓
<div style="display: flex; justify-content: space-between;"> <span><b>I have no phlegm (mucus) in my chest at all</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>My chest is completely full of phlegm (mucus)</b></span> </div>	↓
<div style="display: flex; justify-content: space-between;"> <span><b>My chest does not feel tight at all</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>My chest feels very tight</b></span> </div>	↓
<div style="display: flex; justify-content: space-between;"> <span><b>When I walk up a hill or one flight of stairs I am not breathless</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>When I walk up a hill or one flight of stairs I am very breathless</b></span> </div>	↓
<div style="display: flex; justify-content: space-between;"> <span><b>I am not limited doing any activities at home</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>I am very limited doing activities at home</b></span> </div>	↓
<div style="display: flex; justify-content: space-between;"> <span><b>I am confident leaving my home despite my lung condition</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>I am not at all confident leaving my home because of my lung condition</b></span> </div>	↓
<div style="display: flex; justify-content: space-between;"> <span><b>I sleep soundly</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>I don't sleep soundly because of my lung condition</b></span> </div>	↓
<div style="display: flex; justify-content: space-between;"> <span><b>I have lots of energy</b></span> <span>(0) <input type="radio"/> (1) <input type="radio"/> (2) <input type="radio"/> (3) <input type="radio"/> (4) <input type="radio"/> (5)</span> <span><b>I have no energy at all</b></span> </div>	↓
<b>TOTAL SCORE</b>	↓
	□

COPD Assessment Test and the CAT logo is a trade mark of the GlaxoSmithKline group of companies.  
© 2009 GlaxoSmithKline group of companies. All rights reserved.  
Last Updated: February 24, 2012

Figure C.1: COPD Assessment Test [CAT]

# Appendix D

## List of Features (Data-Driven Approach)

Time in 0-5min Bouts Inactive in % of Total Time  
Time in 0-5min Bouts Inactive in % of Total Time (1 of 3)  
Time in 0-5min Bouts Inactive in % of Total Time (2 of 3)  
Time in 0-5min Bouts Inactive in % of Total Time (3 of 3)  
Time in 0-5min Bouts Light Active in % of Total Time  
Time in 0-5min Bouts Light Active in % of Total Time (1 of 3)  
Time in 0-5min Bouts Light Active in % of Total Time (2 of 3)  
Time in 0-5min Bouts Light Active in % of Total Time (3 of 3)  
Time in 0-5min Bouts High Active in % of Total Time  
Time in 0-5min Bouts High Active in % of Total Time (1 of 3)  
Time in 0-5min Bouts High Active in % of Total Time (2 of 3)  
Time in 0-5min Bouts High Active in % of Total Time (3 of 3)  
Time in 5-10min Bouts Inactive in % of Total Time  
Time in 5-10min Bouts Inactive in % of Total Time (1 of 3)  
Time in 5-10min Bouts Inactive in % of Total Time (2 of 3)  
Time in 5-10min Bouts Inactive in % of Total Time (3 of 3)  
Time in 5-10min Bouts Light Active in % of Total Time  
Time in 5-10min Bouts Light Active in % of Total Time (1 of 3)  
Time in 5-10min Bouts Light Active in % of Total Time (2 of 3)  
Time in 5-10min Bouts Light Active in % of Total Time (3 of 3)  
Time in 5-10min Bouts High Active in % of Total Time  
Time in 5-10min Bouts High Active in % of Total Time (1 of 3)  
Time in 5-10min Bouts High Active in % of Total Time (2 of 3)

Time in 5-10min Bouts High Active in % of Total Time (3 of 3)  
Time in 10-20min Bouts Inactive in % of Total Time  
Time in 10-20min Bouts Inactive in % of Total Time (1 of 3)  
Time in 10-20min Bouts Inactive in % of Total Time (2 of 3)  
Time in 10-20min Bouts Inactive in % of Total Time (3 of 3)  
Time in 10-20min Bouts Light Active in % of Total Time  
Time in 10-20min Bouts Light Active in % of Total Time (1 of 3)  
Time in 10-20min Bouts Light Active in % of Total Time (2 of 3)  
Time in 10-20min Bouts Light Active in % of Total Time (3 of 3)  
Time in 10-20min Bouts High Active in % of Total Time  
Time in 10-20min Bouts High Active in % of Total Time (1 of 3)  
Time in 10-20min Bouts High Active in % of Total Time (2 of 3)  
Time in 10-20min Bouts High Active in % of Total Time (3 of 3)  
Time in 20-30min Bouts Inactive in % of Total Time  
Time in 20-30min Bouts Inactive in % of Total Time (1 of 3)  
Time in 20-30min Bouts Inactive in % of Total Time (2 of 3)  
Time in 20-30min Bouts Inactive in % of Total Time (3 of 3)  
Time in 20-30min Bouts Light Active in % of Total Time  
Time in 20-30min Bouts Light Active in % of Total Time (1 of 3)  
Time in 20-30min Bouts Light Active in % of Total Time (2 of 3)  
Time in 20-30min Bouts Light Active in % of Total Time (3 of 3)  
Time in 20-30min Bouts High Active in % of Total Time  
Time in 20-30min Bouts High Active in % of Total Time (1 of 3)  
Time in 20-30min Bouts High Active in % of Total Time (2 of 3)  
Time in 20-30min Bouts High Active in % of Total Time (3 of 3)  
Time in 30-60min Bouts Inactive in % of Total Time  
Time in 30-60min Bouts Inactive in % of Total Time (1 of 3)  
Time in 30-60min Bouts Inactive in % of Total Time (2 of 3)  
Time in 30-60min Bouts Inactive in % of Total Time (3 of 3)  
Time in 30-60min Bouts Light Active in % of Total Time  
Time in 30-60min Bouts Light Active in % of Total Time (1 of 3)  
Time in 30-60min Bouts Light Active in % of Total Time (2 of 3)  
Time in 30-60min Bouts Light Active in % of Total Time (3 of 3)  
Time in 30-60min Bouts High Active in % of Total Time

Time in 30-60min Bouts High Active in % of Total Time (1 of 3)

Time in 30-60min Bouts High Active in % of Total Time (2 of 3)

Time in 30-60min Bouts High Active in % of Total Time (3 of 3)

Time in > 60min Bouts Inactive in % of Total Time

Time in > 60min Bouts Inactive in % of Total Time (1 of 3)

Time in > 60min Bouts Inactive in % of Total Time (2 of 3)

Time in > 60min Bouts Inactive in % of Total Time (3 of 3)

Time in > 60min Bouts Light Active in % of Total Time

Time in > 60min Bouts Light Active in % of Total Time (1 of 3)

Time in > 60min Bouts Light Active in % of Total Time (2 of 3)

Time in > 60min Bouts Light Active in % of Total Time (3 of 3)

Time in > 60min Bouts High Active in % of Total Time

Time in > 60min Bouts High Active in % of Total Time (1 of 3)

Time in > 60min Bouts High Active in % of Total Time (2 of 3)

Time in > 60min Bouts High Active in % of Total Time (3 of 3)

Information Entropy

Information Entropy - High Active vs. Not High Active

Information Entropy - Inactive vs. Active

Longest Time Sedentary in % of Total Time

Total Mean

Total Mean (1 of 3)

Total Mean (2 of 3)

Total Mean (3 of 3)

Time in 60-120min Bouts Inactive in % of Total Time

Time in 120-180min Bouts Inactive in % of Total Time

Time in 180-240min Bouts Inactive in % of Total Time

Time in 240-300min Bouts Inactive in % of Total Time

Time in 300-360min Bouts Inactive in % of Total Time

Time in 420-480min Bouts Inactive in % of Total Time

Time in 480-540min Bouts Inactive in % of Total Time

Time in 540-600min Bouts Inactive in % of Total Time

Total Time Inactive in % of Total Time

Total Time Light Active in % of Total Time

Total Time High Active in % of Total Time



# List of Figures

2.1	ActiGraph wGT3x-BT [Act09] . . . . .	12
3.1	Example of a barcode with a resolution of 120s . . . . .	15
4.1	The result of the principal component analysis for the data driven approach. The first three principal components displayed here accounted for 29.9% of the total variance. . . . .	24
4.2	The average silhouette scores for a different number of clusters $k$ ( $k \in [2, 20]$ ) of the agglomerative clustering within the data-driven approach . . . . .	26
4.3	Dendrogram of the data-driven approach . . . . .	26
4.4	Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the agglomerative clustering within the data-driven approach on right side ( $k = 2$ ) . . . . .	27
4.5	The agglomerative clustering result for the data-driven approach ( $k = 2$ ) . . . . .	27
4.6	Representative barcodes for Cluster 1 of the agglomerative clustering result within the data-driven approach . . . . .	28
4.7	Representative barcodes for Cluster 2 of the agglomerative clustering result within the data-driven approach . . . . .	28
4.8	The average silhouette scores for a different number of clusters $k$ ( $k \in [2, 20]$ ) of the agglomerative clustering within the data-driven approach . . . . .	30
4.9	Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the k-means within the data-driven approach on the right side ( $k = 2$ ) . . . . .	30
4.10	The k-means clustering result for the data-driven approach ( $k = 2$ ) . . . . .	31
4.11	Representative barcodes for Cluster 1 of the k-means clustering result within the data-driven approach . . . . .	32

4.12	Representative barcodes for Cluster 2 of the k-means clustering result within the data-driven approach . . . . .	32
4.13	The result of the principal component analysis for the data driven approach. The first three principal components accounted for 94.8% of the total variance. . . . .	34
4.14	The average silhouette scores for a different number of clusters $k$ ( $k \in [2, 20]$ ) of the agglomerative clustering within the expert-driven approach . . . . .	36
4.15	Dendrogram of the expert-driven approach . . . . .	36
4.16	Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the agglomerative clustering within the expert-driven approach on the right side ( $k = 2$ ) . . . . .	37
4.17	The agglomerative clustering result for the expert-driven approach ( $k = 2$ ) . . . . .	37
4.18	Representative barcodes for Cluster 1 of the agglomerative clustering result within the expert-driven approach . . . . .	38
4.19	Representative barcodes for Cluster 2 of the agglomerative clustering result within the expert-driven approach . . . . .	38
4.20	The average silhouette scores for a different number of clusters $k$ ( $k \in [2, 20]$ ) of the k-means clustering within the expert-driven approach . . . . .	40
4.21	Silhouette value of every data point with the average silhouette score on the left side and the corresponding cluster result of the agglomerative clustering within the expert-driven approach on the right side ( $k = 4$ ) . . . . .	40
4.22	The k-means clustering result for the expert-driven approach ( $k = 4$ ) . . . . .	41
4.23	Representative barcodes for Cluster 1 of the k-means clustering result within the expert-driven approach . . . . .	42
4.24	Representative barcodes for Cluster 2 of the k-means clustering result within the expert-driven approach . . . . .	42
4.25	Representative barcodes for Cluster 3 of the k-means clustering result within the expert-driven approach . . . . .	43
4.26	Representative barcodes for Cluster 4 of the k-means clustering result within the expert-driven approach . . . . .	43
C.1	COPD Assessment Test [CAT] . . . . .	58

# List of Tables

2.1	Classification of airflow limitation severity in COPD . . . . .	9
2.2	COPD ladder of poor health . . . . .	11
3.1	General characteristics of the COPD patients . . . . .	14
3.2	List of features for the expert-driven approach . . . . .	17
4.1	Average characteristics of the different clusters of the agglomerative clustering algorithm within the data-driven approach . . . . .	25
4.2	Average characteristics of the different clusters of the k-mean algorithm within the data-driven approach . . . . .	29
4.3	Correlation between the original features of the expert-driven approach and the resulting principal components (PC) . . . . .	33
4.4	Average characteristics of the different clusters of the agglomerative algorithm within the expert-driven approach **: Significant difference between Cluster 1 and Cluster 2 ( $p < 0.001$ ) . . . . .	35
4.5	Average characteristics of the different clusters of the agglomerative algorithm within the expert-driven approach **: Significant difference between all clusters ( $p < 0.001$ ) * : Significant difference between all clusters except between Cluster 2 and Cluster 4 ( $p < 0.001$ ) . . . . .	39



# Bibliography

- [Act09] *ActiGraph wGT3x-BT*, [www.actigraphcorp.com/actigraph-wgt3x-bt](http://www.actigraphcorp.com/actigraph-wgt3x-bt), 2009, Accessed: 2019-03-19.
- [Bar91] D. J. Barker, K. Godfrey, C. Fall, C. Osmond, P. Winter, S. Shaheen: *Relation of birth weight and childhood respiratory infection to adult lung function and death from chronic obstructive airways disease.*, *Bmj*, Bd. 303, Nr. 6804, 1991, S. 671–675.
- [Bis06] C. M. Bishop: *Pattern recognition and machine learning*, springer, 2006.
- [Bla95] J. M. Bland, D. G. Altman: *Multiple significance tests: the Bonferroni method*, *Bmj*, Bd. 310, Nr. 6973, 1995, S. 170.
- [Bui07] A. S. Buist, M. A. McBurnie, W. M. Vollmer, S. Gillespie, P. Burney, D. M. Mannino, A. M. Menezes, S. D. Sullivan, T. A. Lee, K. B. Weiss, others: *International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study*, *The Lancet*, Bd. 370, Nr. 9589, 2007, S. 741–750.
- [Bus13] J. B. Bussmann, R. J. van den Berg-Emons: *To total amount of activityâ.. and beyond: perspectives on measuring physical behavior*, *Frontiers in psychology*, Bd. 4, 2013, S. 463.
- [Bus14] A. K. Busby, R. L. Reese, S. R. Simon: *Pulmonary rehabilitation maintenance interventions: a systematic review*, *American journal of health behavior*, Bd. 38, Nr. 3, 2014, S. 321–330.
- [Byr16] B. Byrom, D. A. Rowe: *Measuring free-living physical activity in COPD patients: deriving methodology standards for clinical trials through a review of research studies*, *Contemporary clinical trials*, Bd. 47, 2016, S. 172–184.
- [Cas08] C. Casanova, C. Cote, J. M. Marin, V. Pinto-Plata, J. P. de Torres, A. Aguirre-Jaíme, C. Vassaux, B. R. Celli: *Distance and oxygen desaturation during the 6-min walk test*

- as predictors of long-term mortality in patients with COPD*, *Chest*, Bd. 134, Nr. 4, 2008, S. 746–752.
- [CAT] *COPD Assessment Test*, <https://www.catestonline.org/>, Accessed: 2019-03-19.
- [Cha10] S. F. Chastin, K. Baker, D. Jones, D. Burn, M. H. Granat, L. Rochester: *The pattern of habitual sedentary behavior is different in advanced Parkinson's disease*, *Movement Disorders*, Bd. 25, Nr. 13, 2010, S. 2114–2120.
- [Cho10] M. H. Cho, N. Boutaoui, B. J. Klanderman, J. S. Sylvia, J. P. Ziniti, C. P. Hersh, D. L. DeMeo, G. M. Hunninghake, A. A. Litonjua, D. Sparrow, others: *Variants in FAM13A are associated with chronic obstructive pulmonary disease*, *Nature genetics*, Bd. 42, Nr. 3, 2010, S. 200.
- [Cho14] M. H. Cho, M.-L. N. McDonald, X. Zhou, M. Mattheisen, P. J. Castaldi, C. P. Hersh, D. L. DeMeo, J. S. Sylvia, J. Ziniti, N. M. Laird, others: *Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis*, *The lancet Respiratory medicine*, Bd. 2, Nr. 3, 2014, S. 214–225.
- [CN12] L. W. Cindy Ng, J. Mackney, S. Jenkins, K. Hill: *Does exercise training change physical activity in people with COPD? A systematic review and meta-analysis*, *Chronic respiratory disease*, Bd. 9, Nr. 1, 2012, S. 17–26.
- [Edw93] L. Edwards: *Applied analysis of variance in behavioral science*, Bd. 137, CRC Press, 1993.
- [Ega12] C. Egan, B. M. Deering, C. Blake, B. M. Fullen, N. M. McCormack, M. A. Spruit, R. W. Costello: *Short term and long term effects of pulmonary rehabilitation on physical activity in COPD*, *Respiratory medicine*, Bd. 106, Nr. 12, 2012, S. 1671–1679.
- [Eis10] M. D. Eisner, N. Anthonisen, D. Coultas, N. Kuenzli, R. Perez-Padilla, D. Postma, I. Romieu, E. K. Silverman, J. R. Balmes: *An official American Thoracic Society public policy statement: Novel risk factors and the global burden of chronic obstructive pulmonary disease*, *American journal of respiratory and critical care medicine*, Bd. 182, Nr. 5, 2010, S. 693–718.
- [Eve15] K. R. Evenson, F. Wen, J. S. Metzger, A. H. Herring: *Physical activity and sedentary behavior patterns using accelerometry from a national sample of United States adults*,

*International Journal of Behavioral Nutrition and Physical Activity*, Bd. 12, Nr. 1, 2015, S. 20.

- [Fer09] C. J. Ferguson: *An effect size primer: A guide for clinicians and researchers.*, *Professional Psychology: Research and Practice*, Bd. 40, Nr. 5, 2009, S. 532.
- [Fog99] K. Foglio, L. Bianchi, G. Bruletti, L. Battista, M. Pagani, N. Ambrosino: *Long-term effectiveness of pulmonary rehabilitation in patients with chronic airway obstruction*, *European Respiratory Journal*, Bd. 13, Nr. 1, 1999, S. 125–132.
- [Fre98] P. S. Freedson, E. Melanson, J. Sirard: *Calibration of the Computer Science and Applications, Inc. accelerometer.*, *Medicine and science in sports and exercise*, Bd. 30, Nr. 5, 1998, S. 777–781.
- [GA06] J. Garcia-Aymerich, P. Lange, M. Benet, P. Schnohr, J. M. Antó: *Regular physical activity reduces hospital admission and mortality in chronic obstructive pulmonary disease: a population based cohort study*, *Thorax*, Bd. 61, Nr. 9, 2006, S. 772–778.
- [GA07] J. Garcia-Aymerich, P. Lange, M. Benet, P. Schnohr, J. M. Antó: *Regular physical activity modifies smoking-related lung function decline and reduces risk of chronic obstructive pulmonary disease: a population-based cohort study*, *American journal of respiratory and critical care medicine*, Bd. 175, Nr. 5, 2007, S. 458–463.
- [Gei17] W. Geidl, J. Semrau, R. Streber, N. Lehbort, S. Wingart, A. Tallner, M. Wittmann, R. Wagner, K. Schultz, K. Pfeifer: *Effects of a brief, pedometer-based behavioral intervention for individuals with COPD during inpatient pulmonary rehabilitation on 6-week and 6-month objectively measured physical activity: study protocol for a randomized controlled trial*, *Trials*, Bd. 18, Nr. 1, 2017, S. 396.
- [Glo18] R. Gloeckl, T. Schneeberger, I. Jarosch, K. Kenn: *Pulmonary rehabilitation and exercise training in chronic obstructive pulmonary disease*, *Deutsches Ärzteblatt International*, Bd. 115, Nr. 8, 2018, S. 117.
- [GS14] E. Gimeno-Santos, A. Frei, C. Steurer-Stey, J. De Batlle, R. A. Rabinovich, Y. Raste, N. S. Hopkinson, M. I. Polkey, H. Van Remoortel, T. Troosters, others: *Determinants and outcomes of physical activity in patients with COPD: a systematic review*, *Thorax*, Bd. 69, Nr. 8, 2014, S. 731–739.

- [Hal06] R. Halbert, J. Natoli, A. Gano, E. Badamgarav, A. S. Buist, D. Mannino: *Global burden of COPD: systematic review and meta-analysis*, *European Respiratory Journal*, Bd. 28, Nr. 3, 2006, S. 523–532.
- [Hea11] G. N. Healy, C. E. Matthews, D. W. Dunstan, E. A. Winkler, N. Owen: *Sedentary time and cardio-metabolic biomarkers in US adults: NHANES 2003–06*, *European heart journal*, Bd. 32, Nr. 5, 2011, S. 590–597.
- [Hil12] K. Hill, T. E. Dolmage, L. Woon, D. Coutts, R. Goldstein, D. Brooks: *Defining the relationship between average daily energy expenditure and field-based walking tests and aerobic reserve in COPD*, *Chest*, Bd. 141, Nr. 2, 2012, S. 406–412.
- [Jai99] A. K. Jain, M. N. Murty, P. J. Flynn: *Data clustering: a review*, *ACM computing surveys (CSUR)*, Bd. 31, Nr. 3, 1999, S. 264–323.
- [Joh12] D. John, P. Freedson: *ActiGraph and Actical physical activity monitors: a peek under the hood*, *Medicine and science in sports and exercise*, Bd. 44, Nr. 1 Suppl 1, 2012, S. S86.
- [Jol11] I. Jolliffe: *Principal component analysis*, Springer, 2011.
- [Jon09a] P. W. Jones: *Health status and the spiral of decline*, *COPD: Journal of Chronic Obstructive Pulmonary Disease*, Bd. 6, Nr. 1, 2009, S. 59–63.
- [Jon09b] P. Jones, G. Harding, P. Berry, I. Wiklund, W. Chen, N. K. Leidy: *Development and first validation of the COPD Assessment Test*, *European Respiratory Journal*, Bd. 34, Nr. 3, 2009, S. 648–654.
- [Jon11] P. W. Jones, M. Tabberer, W.-H. Chen: *Creating scenarios of the impact of COPD and their relationship to COPD Assessment Test (CAT) scores*, *BMC pulmonary medicine*, Bd. 11, Nr. 1, 2011, S. 42.
- [Jus81] B. Justusson: *Median filtering: Statistical properties*, in *Two-Dimensional Digital Signal Processing II*, Springer, 1981, S. 161–196.
- [Kri09] H.-P. Kriegel, P. Kröger, A. Zimek: *Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering*, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Bd. 3, Nr. 1, 2009, S. 1.

- [Lan15] P. Lange, B. Celli, A. Agustí, G. Boje Jensen, M. Divo, R. Faner, S. Guerra, J. L. Marott, F. D. Martinez, P. Martinez-Camblor, others: *Lung-function trajectories leading to chronic obstructive pulmonary disease*, *New England Journal of Medicine*, Bd. 373, Nr. 2, 2015, S. 111–122.
- [Lee13] P. H. Lee, Y.-Y. Yu, I. McDowell, G. M. Leung, T. Lam: *A cluster analysis of patterns of objectively measured physical activity in Hong Kong*, *Public health nutrition*, Bd. 16, Nr. 8, 2013, S. 1436–1444.
- [Loz12] R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn, others: *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010*, *The lancet*, Bd. 380, Nr. 9859, 2012, S. 2095–2128.
- [Mac67] J. MacQueen, others: *Some methods for classification and analysis of multivariate observations*, in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Bd. 1, Oakland, CA, USA, 1967, S. 281–297.
- [Man16] L. C. Mantoani, N. Rubio, B. McKinstry, W. MacNee, R. A. Rabinovich: *Interventions to modify physical activity in patients with COPD: a systematic review*, *European Respiratory Journal*, Bd. 48, Nr. 1, 2016, S. 69–81.
- [Mat06] C. D. Mathers, D. Loncar: *Projections of global mortality and burden of disease from 2002 to 2030*, *PLoS medicine*, Bd. 3, Nr. 11, 2006, S. e442.
- [McC01] S. C. McCLOSKEY, B. D. Patel, S. J. Hinchliffe, E. D. Reid, N. J. Wareham, D. A. Lomas: *Siblings of patients with severe chronic obstructive pulmonary disease have a significant risk of airflow obstruction*, *American journal of respiratory and critical care medicine*, Bd. 164, Nr. 8, 2001, S. 1419–1424.
- [McC15] B. McCarthy, D. Casey, D. Devane, K. Murphy, E. Murphy, Y. Lacasse: *Pulmonary rehabilitation for chronic obstructive pulmonary disease*, *Cochrane database of systematic reviews*, , Nr. 2, 2015.
- [McV16] J. A. McVeigh, E. A. Winkler, E. K. Howie, M. S. Tremblay, A. Smith, R. A. Abbott, P. R. Eastwood, G. N. Healy, L. M. Straker: *Objectively measured patterns of sedentary time and physical activity in young adults of the Raine study cohort*, *International Journal of Behavioral Nutrition and Physical Activity*, Bd. 13, Nr. 1, 2016, S. 41.

- [Mer15] N. Mercado, K. Ito, P. J. Barnes: *Accelerated ageing of the lung in COPD: new concepts*, *Thorax*, Bd. 70, Nr. 5, 2015, S. 482–489.
- [Mes17] R. Mesquita, G. Spina, F. Pitta, D. Donaire-Gonzalez, B. M. Deering, M. S. Patel, K. E. Mitchell, J. Alison, A. J. Van Gestel, S. Zogg, others: *Physical activity patterns and clusters in 1001 patients with COPD*, *Chronic respiratory disease*, Bd. 14, Nr. 3, 2017, S. 256–269.
- [Mir14] M. Miravittles, H. Worth, J. J. S. Cataluña, D. Price, F. De Benedetto, N. Roche, N. S. Godtfredsen, T. van Der Molen, C.-G. Löfdahl, L. Padullés, others: *Observational study to characterise 24-hour COPD symptoms and their relationship with patient-reported outcomes: results from the ASSESS study*, *Respiratory research*, Bd. 15, Nr. 1, 2014, S. 122.
- [Mur12] C. J. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, others: *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*, Bd. 380, Nr. 9859, 2012, S. 2197–2223.
- [Pat93] S. M. Patterson, D. S. Krantz, L. C. Montgomery, P. A. Deuster, S. M. Hedges, L. E. Nebel: *Automated physical activity monitoring: Validation and comparison with physiological and self-report measures*, *Psychophysiology*, Bd. 30, Nr. 3, 1993, S. 296–305.
- [Ped11] 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, E. Duchesnay: *Scikit-learn: Machine Learning in Python*, *Journal of Machine Learning Research*, Bd. 12, 2011, S. 2825–2830.
- [PI12] A. Paraschiv-Ionescu, C. Perruchoud, E. Buchser, K. Aminian: *Barcoding human physical activity to assess chronic pain conditions*, *PloS one*, Bd. 7, Nr. 2, 2012, S. e32239.
- [Pil09] S. G. Pillai, D. Ge, G. Zhu, X. Kong, K. V. Shianna, A. C. Need, S. Feng, C. P. Hersh, P. Bakke, A. Gulsvik, others: *A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci*, *PLoS genetics*, Bd. 5, Nr. 3, 2009, S. e1000421.

- [Pit05] F. Pitta, T. Troosters, M. A. Spruit, V. S. Probst, M. Decramer, R. Gosselink: *Characteristics of physical activities in daily life in chronic obstructive pulmonary disease*, *American journal of respiratory and critical care medicine*, Bd. 171, Nr. 9, 2005, S. 972–977.
- [Pit06a] F. Pitta, T. Troosters, V. Probst, M. Spruit, M. Decramer, R. Gosselink: *Quantifying physical activity in daily life with questionnaires and motion sensors in COPD*, *European respiratory journal*, Bd. 27, Nr. 5, 2006, S. 1040–1055.
- [Pit06b] F. Pitta, T. Troosters, V. S. Probst, S. Lucas, M. Decramer, R. Gosselink: *Potential consequences for stable chronic obstructive pulmonary disease patients who do not get the recommended minimum daily amount of physical activity*, *Jornal Brasileiro de Pneumologia*, Bd. 32, Nr. 4, 2006, S. 301–308.
- [Pit08] F. Pitta, M. Y. Takaki, N. H. de Oliveira, T. J. Sant’Anna, A. D. Fontana, D. Kovelis, C. A. Camillo, V. S. Probst, A. F. Brunetto: *Relationship between pulmonary function and physical activity in daily life in patients with COPD*, *Respiratory medicine*, Bd. 102, Nr. 8, 2008, S. 1203–1207.
- [Ren06] S. I. Rennard, J. Vestbo: *COPD: the dangerous underestimate of 15%*, *The Lancet*, Bd. 367, Nr. 9518, 2006, S. 1216–1219.
- [Ren11] E. Rendón, I. Abundez, A. Arizmendi, E. M. Quiroz: *Internal versus external cluster validation indexes*, *International Journal of computers and communications*, Bd. 5, Nr. 1, 2011, S. 27–34.
- [Rep10] E. Repapi, I. Sayers, L. V. Wain, P. R. Burton, T. Johnson, M. Obeidat, J. H. Zhao, A. Ramasamy, G. Zhai, V. Vitart, others: *Genome-wide association study identifies five loci associated with lung function*, *Nature genetics*, Bd. 42, Nr. 1, 2010, S. 36.
- [Rie07] A. L. Ries, G. S. Bauldoff, B. W. Carlin, R. Casaburi, C. F. Emery, D. A. Mahler, B. Make, C. L. Rochester, R. ZuWallack, C. Herrerias: *Pulmonary rehabilitation: joint ACCP/AACVPR evidence-based clinical practice guidelines*, *Chest*, Bd. 131, Nr. 5, 2007, S. 4S–42S.
- [Roc06] L. Rochester, D. Jones, V. Hetherington, A. Nieuwboer, A.-M. Willems, G. Kwakkel, E. V. Wegen: *Gait and gait-related activities and fatigue in Parkinson’s disease: what is the relationship?*, *Disability and rehabilitation*, Bd. 28, Nr. 22, 2006, S. 1365–1371.

- [Roc15] C. L. Rochester, I. Vogiatzis, A. E. Holland, S. C. Lareau, D. D. Marciniuk, M. A. Puhan, M. A. Spruit, S. Masefield, R. Casaburi, E. M. Clini, others: *An official American Thoracic Society/European Respiratory Society policy statement: enhancing implementation, use, and delivery of pulmonary rehabilitation*, *American journal of respiratory and critical care medicine*, Bd. 192, Nr. 11, 2015, S. 1373–1386.
- [Rok05] L. Rokach, O. Maimon: *Clustering methods*, in *Data mining and knowledge discovery handbook*, Springer, 2005, S. 321–352.
- [Rou87] P. J. Rousseeuw: *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, *Journal of computational and applied mathematics*, Bd. 20, 1987, S. 53–65.
- [SA11] M. Soler Artigas, L. V. Wain, E. Repapi, M. Obeidat, I. Sayers, P. R. Burton, T. Johnson, J. H. Zhao, E. Albrecht, A. F. Dominiczak, others: *Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function*, *American journal of respiratory and critical care medicine*, Bd. 184, Nr. 7, 2011, S. 786–795.
- [Sah16] H. Sahin, I. Naz, Y. Varol, N. Aksel, F. Tuksavul, A. Ozsoz: *Is a pulmonary rehabilitation program effective in COPD patients with chronic hypercapnic failure?*, *Expert review of respiratory medicine*, Bd. 10, Nr. 5, 2016, S. 593–598.
- [Sal09] S. S. Salvi, P. J. Barnes: *Chronic obstructive pulmonary disease in non-smokers*, *The lancet*, Bd. 374, Nr. 9691, 2009, S. 733–743.
- [Sas11] J. E. Sasaki, D. John, P. S. Freedson: *Validation and comparison of ActiGraph activity monitors*, *Journal of Science and Medicine in Sport*, Bd. 14, Nr. 5, 2011, S. 411–416.
- [Sha48] C. E. Shannon: *A mathematical theory of communication*, *Bell system technical journal*, Bd. 27, Nr. 3, 1948, S. 379–423.
- [SL12] A. Santos-Lozano, P. J. Marín, G. Torres-Luque, J. R. Ruiz, A. Lucía, N. Garatachea: *Technical variability of the GT3X accelerometer*, *Medical engineering & physics*, Bd. 34, Nr. 6, 2012, S. 787–790.
- [Spr13] M. A. Spruit, S. J. Singh, C. Garvey, R. ZuWallack, L. Nici, C. Rochester, K. Hill, A. E. Holland, S. C. Lareau, W. D.-C. Man, others: *An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary*

*rehabilitation, American journal of respiratory and critical care medicine*, Bd. 188, Nr. 8, 2013, S. e13–e64.

- [Spr15] M. A. Spruit, F. Pitta, E. McAuley, R. L. ZuWallack, L. Nici: *Pulmonary rehabilitation and physical activity in patients with chronic obstructive pulmonary disease, American journal of respiratory and critical care medicine*, Bd. 192, Nr. 8, 2015, S. 924–933.
- [Ste03] B. G. Steele, B. Belza, J. Hunziker, L. Holt, M. Legro, J. Coppersmith, D. Buchner, S. Lakshminaryan: *Monitoring daily activity during pulmonary rehabilitation using a triaxial accelerometer, Journal of Cardiopulmonary Rehabilitation and Prevention*, Bd. 23, Nr. 2, 2003, S. 139–142.
- [Ste07] D. A. Stern, W. J. Morgan, A. L. Wright, S. Guerra, F. D. Martinez: *Poor airway function in early infancy and lung function by age 22 years: a non-selective longitudinal cohort study, The Lancet*, Bd. 370, Nr. 9589, 2007, S. 758–764.
- [Sto05] J. K. Stoller, L. S. Aboussouan:  *$\alpha$ 1-antitrypsin deficiency, The Lancet*, Bd. 365, Nr. 9478, 2005, S. 2225–2236.
- [Str93] G. Strang, G. Strang, G. Strang, G. Strang: *Introduction to linear algebra*, Bd. 3, Wellesley-Cambridge Press Wellesley, MA, 1993.
- [Swa00] A. M. Swartz, S. J. Strath, D. R. BASSETT, W. L. OâBRIEN, G. A. King, B. E. Ainsworth: *Estimation of energy expenditure using CSA accelerometers at hip and wrist sites, Medicine & Science in Sports & Exercise*, Bd. 32, Nr. 9, 2000, S. S450–S456.
- [Tas92] D. P. Tashkin, M. D. Altose, E. R. Bleecker, J. E. Connett, R. E. Kanner, W. W. Lee, R. Wise: *The Lung Health Study: airway responsiveness to inhaled methacholine in smokers with mild to moderate airflow limitation., American Journal of Respiratory and Critical Care Medicine*, Bd. 145, Nr. 2, 1992, S. 301–310.
- [Tod93] T. Todisco, F. De Benedictis, L. Iannacci, S. Baglioni, A. Eslami, E. Todisco, M. Dottorini: *Mild prematurity and respiratory functions, European journal of pediatrics*, Bd. 152, Nr. 1, 1993, S. 55–58.
- [Tro00] T. Troosters, R. Gosselink, M. Decramer: *Short-and long-term effects of outpatient rehabilitation in patients with chronic obstructive pulmonary disease: a randomized trial, The American journal of medicine*, Bd. 109, Nr. 3, 2000, S. 207–212.

- [Try96] W. W. Tryon, R. Williams: *Fully proportional actigraphy: a new instrument*, *Behavior Research Methods, Instruments, & Computers*, Bd. 28, Nr. 3, 1996, S. 392–403.
- [vB18] A. R. van Buul, M. J. Kasteleyn, N. H. Chavannes, C. Taube: *Physical activity in the morning and afternoon is lower in patients with chronic obstructive pulmonary disease with morning symptoms*, *Respiratory research*, Bd. 19, Nr. 1, 2018, S. 49.
- [Vog17] C. F. Vogelmeier, G. J. Criner, F. J. Martinez, A. Anzueto, P. J. Barnes, J. Bourbeau, B. R. Celli, R. Chen, M. Decramer, L. M. Fabbri, others: *Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary*, *American journal of respiratory and critical care medicine*, Bd. 195, Nr. 5, 2017, S. 557–582.
- [War06] D. E. Warburton, C. W. Nicol, S. S. Bredin: *Health benefits of physical activity: the evidence*, *Cmaj*, Bd. 174, Nr. 6, 2006, S. 801–809.
- [Wat09] H. Watz, B. Waschki, T. Meyer, H. Magnussen: *Physical activity in patients with COPD*, *European Respiratory Journal*, Bd. 33, Nr. 2, 2009, S. 262–272.
- [Wel00] G. J. Welk, S. N. Blair, K. Wood, S. Jones, R. W. Thompson: *A comparative evaluation of three accelerometry-based physical activity monitors*, *Medicine & Science in Sports & Exercise*, Bd. 32, Nr. 9, 2000, S. S489–S497.
- [Wel02] G. Welk: *Physical activity assessments for health-related research*, Human Kinetics, 2002.
- [WJ63] J. H. Ward Jr: *Hierarchical grouping to optimize an objective function*, *Journal of the American statistical association*, Bd. 58, Nr. 301, 1963, S. 236–244.
- [Yaw09] B. Yawn, D. Mannino, T. Littlejohn, G. Ruoff, A. Emmett, I. Raphiou, G. Crater: *Prevalence of COPD among symptomatic patients in a primary care setting*, *Current medical research and opinion*, Bd. 25, Nr. 11, 2009, S. 2671–2677.
- [Zha08] Z. Zhang, J. Zhang, H. Xue: *Improved K-means clustering algorithm*, in *2008 Congress on Image and Signal Processing*, Bd. 5, IEEE, 2008, S. 169–172.