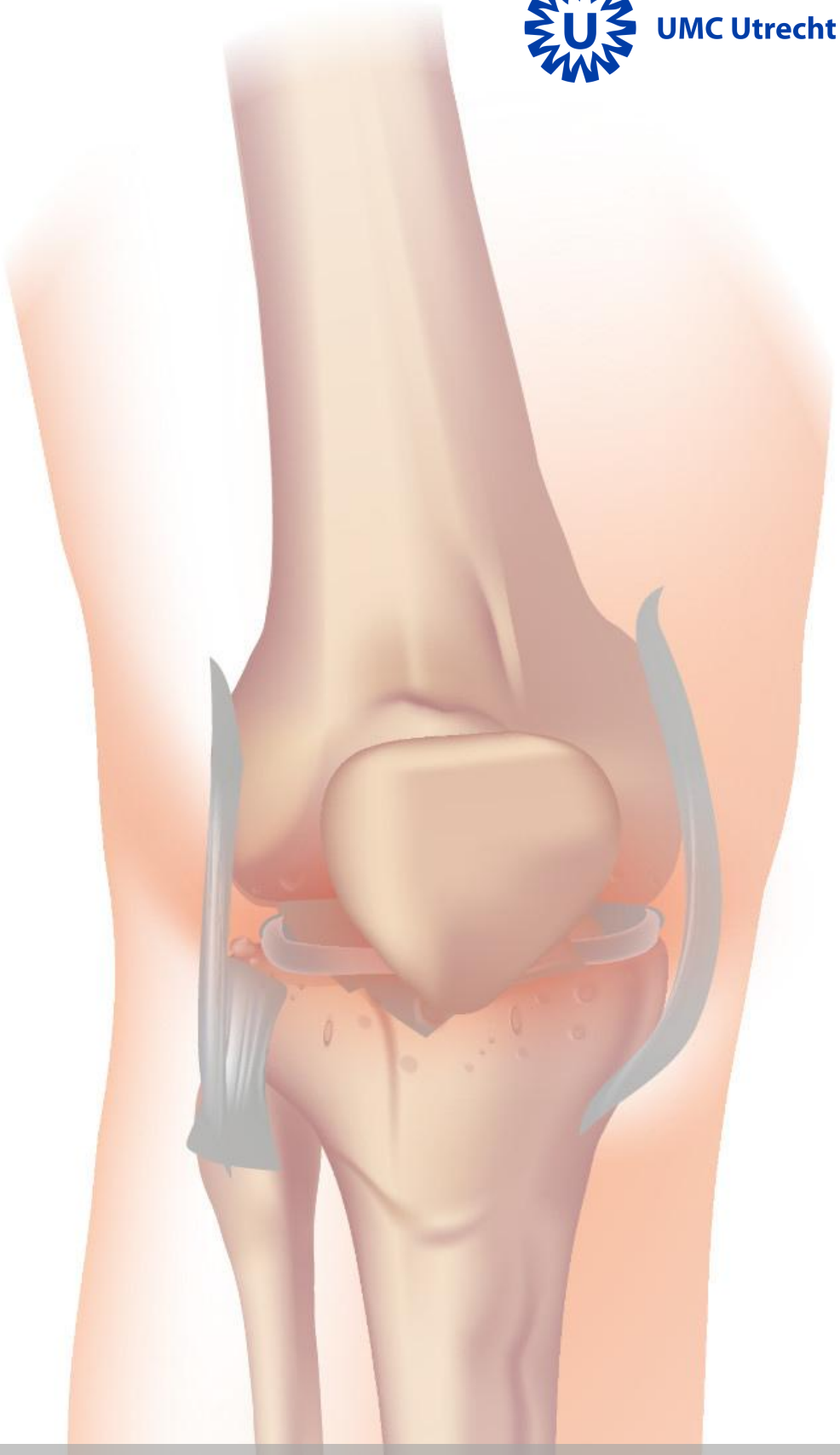
Utrecht University

UMC Utrecht



APPLIED MODEL-BASED CLUSTERING OF FUNCTIONAL DATA - DISTINGUISHING BETWEEN PHENOTYPES OF EARLY KNEE OSTEOARTHRITIS

SARA ALTAMIRANO

UTRECHT UNIVERSITY

MASTER'S THESIS

Applied Model-Based Clustering of Functional Data - Distinguishing Between Phenotypes of Early Knee Osteoarthritis

---

Author:
Sara Altamirano
6485871

Supervisors:
Prof. dr. Y. Velegrakis
dr. Ing. habil G. Krempl

External Supervisors:
dr. P.M.J. Welsing
dr. W.E. van Spil

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in

Business Informatics
Department of Information and Computing Sciences

August 19, 2020

"*Science can be a beautiful tool for discovery, only if it is allowed to dispassionately acknowledge when a more complete picture is emerging.*"

Kelly Brogan, MD

UTRECHT UNIVERSITY

Abstract

Faculty of Science
Department of Information and Computing Sciences

Master of Science

Applied Model-Based Clustering of Functional Data - Distinguishing Between Phenotypes of Early Knee Osteoarthritis

by Sara Altamirano

Context: OA is ranked as the 11th highest contributor to global disability and its prevalence is increasing. By gaining a better understanding of OA heterogeneity, we can potentially contribute to the design of clinical trials, prevention strategies, and treatments. To the best of our knowledge, MBCFD has not been attempted to derive knee OA phenotypes. MBCFD treats the data as curves which can potentially allow us to see complex trends not detected by traditional distance-based clustering algorithms.
Objective: The study aimed to improve OA heterogeneity understanding by testing an MBCFD method's ability to derive clinically-relevant and statistically-significant phenotypes and to assess its performance vis-a-vis a method widely used in the scientific literature.
Methods: This work is based on the CRISP-IDM method. We identified widely-used algorithms in the literature to derive knee OA phenotypes, as well as their characteristics and derived phenotypes. We selected an appropriate MBCFD algorithm and, through iterative data exploration steps with domain experts, we identified clinically-relevant phenotypes with MBCFD as well as computed the statistical significance between the groups. Subsequently, we compared the performance of the MBCFD method to HCA.
Results: MBCFD was able to detect clinically-relevant and statistically-significant knee OA phenotypes for the univariate case. However, for the multivariate case, the phenotypes were clinically relevant but no statistical significance was found between the groups. In addition, MBCFD outperforms HCA in the univariate case but not in the multivariate case.

Acknowledgements

Eight months ago, I embarked on a journey that required me to join two fields: osteoarthritis and machine learning. I knew it was going to be challenging, but I did not realize that by helping to create phenotypes I could potentially influence the future care of a disease that affects millions of people. I am very happy to have gone through this challenging and fulfilling journey, as what I have learned has been life-changing.

This study would not have been possible without several crucial people. Yannis Velegrakis, for his invaluable feedback, patience, understanding, and flexibility. Paco Welsing, for his endurance of my endless questions as well as the daily support and valuable osteoarthritis wisdom. Erwin van Spil, for giving me the opportunity to participate in this project and his positive reinforcement. Georg Krempl, for his valuable insights and kindness. Yvonne Tromp, for providing emotional support and a helping hand.

I would also like to thank the UMCU for facilitating the work, the opportunity, and the means to conduct the study.

I thank my friends and family, who encouraged and supported me throughout. Last, but not least, I would like to thank the loves of my life: my partner and my son, who are the absolute foundation of my existence.

- Sara Altamirano

Contents

List of Figures

List of Tables

Introduction

Osteoarthritis (OA) is the most common form of arthritis; it is a heterogeneous disease characterized by multi-tissue failure in joints and the knee is among the most affected joints (Murphy et al., 2011). Its prevalence is increasing due to the aging population as well as increasingly widespread risk factors, especially obesity and a sedentary lifestyle. Risk factors of OA can be divided into individual factors such as age, gender and diet, and joint-level factors such as injury and malalignment. All risk factors can interact in a complex manner contributing to the heterogeneity of the disease (Palazzo et al., 2016). According to the Global Burden of Disease 2010 study, hip and knee OA were ranked as the 11th highest contributor to global disability (Cross et al., 2014). Vos et al. (2017) estimates that more than 300 million people around the world suffer from OA and Prieto-Alhambra et al. (2014) investigated over three million people from Catalonia, Spain and reported incidence rates of clinically defined OA of 6.5, 2.1, and 2.4/1000 person-years for knee, hip, and hand, respectively.

In 2019, the United Nations announced that the world population could grow to around 8.5 billion in 2030, 9.7 billion in 2050, and 10.9 billion in 2100. In 2018, for the first time in history, persons aged 65 years or over worldwide outnumbered children under age five. Projections indicate that by 2050 there will be more than twice as many persons above 65 as children under five. Overall, the world's population is growing older due to increasing life expectancy and falling fertility levels (United Nations, Department of Economic and Social Affairs, Population Division, 2019). These figures combined potentially represent millions of new symptomatic OA patients as the population grows older, which may be an underestimate as lifestyles, environments, and comorbidities keep changing, and prevalence of obesity is still increasing.

As evidenced by the aforementioned figures, OA is a considerable burden on the world's economy and the healthcare system. In the United States alone, direct and indirect annual costs for OA have been estimated to be over 98 billion US dollars (Brown et al., 2006). In addition, the indirect cost of absenteeism was estimated at approximately 10.3 billion US dollars (Kotlarz et al., 2010). In Spain, the costs of knee and hip OA were of more than 4.7 billion euros in 2007, comparable to 0.5% of the Gross National Product that year (Loza et al., 2009).

Knee OA is an increasingly prevalent joint disease in the Netherlands as well. In 2018, there were nearly 1.5 million people with an OA diagnosis from their general practitioners (GP). Approximately, 65% were female patients and 47% were identified as having OA in the knee. In 2017, expenditure on care for OA amounted to 1.2 billion euros (*Osteoarthritis* 2019).

In recent years, the concept of OA heterogeneity has been gaining acceptance whereas etiological mechanisms are thought to vary between people with OA, resulting in differences in clinical presentation and disease course (Van Spil et al., 2020). OA disease heterogeneity may be most perceptible and relevant in the beginning stages of OA as different etiologic processes may accumulate and coalesce over time in patients (Driban et al., 2010). Moreover, this early phase of the disease is likely also the most opportune moment to achieve life-altering modification of the disease course when symptoms are more manageable and less evident. Additionally, the causes of the disease may be easier to identify in early stages as complications arise over time and symptomatology forms an intertwined network increasingly more difficult to untangle.

OA is a heterogeneous condition described by a variety of clinical features, bio chemical markers, radiographic findings, and different outcomes. No disease-modifying treatment for OA exists, and it is thought that personalization of treatment is required to optimize interventions specifically targeted at subgroups of patients. The main symptoms of knee OA are pain,

stiffness, and loss of function, leading to reduced mobility and quality of life. In general, people start experiencing OA symptoms at a median age of 55 years and live 26 years with the condition (Losina et al., 2013).

The disease course of OA is typically slow, but varies between people and is largely unpredictable (Van Spil et al., 2012). Treatment for both knee and hip OA is symptomatic and usually moderately effective, as no solution influencing the disease course is currently available. Consequently, knee OA is one of the main indications for knee replacement surgery (Conaghan et al., 2010).

In general, treatment recommendations for OA are designed to provide patients with relief of their symptoms, mostly pain. Regrettably, not all patients with OA react uniformly to pain treatments, resulting in the implementation of a mix of interventions aiming for relief which increases the likelihood of adverse health effects and the cost burden on the patient and healthcare system (Kovac et al., 2008).

Although the exact cause of OA remains poorly understood, a number of likely relevant pathobiological and pain mechanisms have been determined (Wesseling et al., 2009). The relevance of these mechanisms might vary between patients because distinct phenotypes "share distinct underlying pathobiological and pain mechanisms and their structural and functional consequences" may exist (Van Spil et al., 2020, p. 1). As mentioned above, the concept of OA heterogeneity has been gaining renewed interest recently in the pursuit of disease-modifying treatment options. Indeed, there are no effective disease-modifying drugs for knee OA, in large part because clinical trials have treated all knee OA as the same disease, disregarding etiology or risk factors (Nelson et al., 2019).

Until recently, it has been asserted that OA patients' *specific* attributes are not clinically considered and treatment options are limited with scarce evidence, particularly for patient-specific interventions (Deveza and Loeser, 2018). Latterly, the development of a number of novel potential treatments with diverse mechanisms of actions has reignited the need to define *homogeneous groups* to personalize treatment given predetermined OA phenotypes and achieve better treatment outcomes (Tonge, Pearson, and Jones, 2014; Karsdal et al., 2016). Appropriately, it has been suggested that tailoring interventions for subgroups of patients, the goal of Precision Medicine, could increase the positive outcomes of treatment and find a more cost-effective solution with less adverse effects (Bruyere et al., 2015; Deveza et al., 2017; Driban et al., 2010; Van Spil, 2012).

Consequently, a subgroup of patients with similar OA characteristics can represent an OA phenotype. In general, a phenotype has been defined as a set of observable characteristics of a subject resulting from environmental and genetic factors. In the medical field, *prognostic phenotyping* is the identification of subgroups that are more likely to reach a specific outcome within a determined period. Identifying prognostic phenotypes is a crucial aspect of designing personalized OA treatments. The optimal manner in which to create phenotypes of OA patients and their clinical value is still under active investigation and is fundamental for the advancement of OA research (Deveza, Nelson, and Loeser, 2019). It is hypothesized that since etiological mechanisms are less entangled early in the disease course (e.g., symptoms of knee OA can begin to appear two or three years before the first instance of radiographic OA), phenotypes could be predicted early in the disease course (Whittle et al., 2016).

Electronic Health Records (EHR) are becoming increasingly more popular and researchers are turning to machine learning methods to be able to mine large amounts of data. Machine learning methods are a new addition to the OA phenotyping field; a variety of methods have been applied to medical data in the last few years for the purpose of identifying phenotypes, most commonly hierarchical cluster analysis (HCA), k-means, latent class analysis (LCA) and

logistic regression (Deveza et al., 2017). High-quality datasets are required to obtain meaningful results. When referring to machine learning, we use the words method and algorithm interchangeably throughout.

Finding meaningful groups in longitudinal data is an unsupervised learning problem because these data do not contain labels (e.g., etiologic mechanisms and progression subgroups are not observed) and previously undetected patterns are being explored. In addition, longitudinal data can essentially be treated as time series or as functions. Clustering functional data is a difficult task because the data fundamentally live in an infinite dimensional space which represents a challenge for traditional multivariate statistics. We used the CHECK (Cohort Hip and Cohort Knee) dataset (Wesseling et al., 2014) for our analysis. Since CHECK data were collected longitudinally over ten years, we believe the time continuum should be considered in the analysis. Data from CHECK are available upon request to all researchers at `http://check-onderzoek.nl`.

The precise issue this pioneering research addressed was to investigate whether functional data analysis (FDA) can contribute to derive clinically-relevant and statistically-significant knee OA phenotypes, specifically by using a model-based clustering for functional data (MBCFD) method. Additionally, we compared MBCFD results to a traditional clustering method (which ignores time) widely used in the literature for knee OA phenotyping. A detailed discussion of the problem statement can be found in chapter 2. This research contributes to OA science since currently we do not have generally-accepted measures to subtype the disease which is deemed highly importantfor developing disease-modifying OA interventions (Deveza and Loeser, 2018).

Research Approach

Phenotypes are observable characteristics of an organism above the molecular level, distinguished by direct observation or finer methods (Johannsen, 1911). If phenotypes are detected or predicted early in the disease course and these are related to the etiological mechanisms of disease this would allow medical practitioners to personalize treatment, design more specific trials, and recommend tailored prevention strategies. Effective strategies to prevent or treat OA are actively anticipated to help give an answer to OA patients as they would aid in lessening the impact of the disease at the individual level (e.g. pain and disability) and at the societal level (e.g. direct and indirect healthcare costs and productivity). When identifying phenotypes with medical records, subgroups are created by clustering similar patients together. Moreover, when using longitudinal data, the patients' trajectories can be taken into account and this is particularly useful for a progressive condition.

Finding phenotypes of early OA in the CHECK dataset is a data mining problem. Thus, the CRISP-IDM method developed by Menger et al. (2016) has been selected for this investigation. CRISP-IDM is an adaptation of the CRISP-DM method (Chapman et al., 2000), applicable to exploratory, iterative, and interactive data analysis in healthcare. Moreover, MBCFD techniques are appropriate for longitudinal, high-dimensional, heterogeneous data as is the case with the different clinical and radiographic assessments made over time for OA patients within the CHECK cohort. In order to answer our research questions (detailed below), first we conducted a literature review to understand the current state of the art, specifically which methods are being used, which characteristics have been found relevant and which MBCFD methods are available for analysis. The method selection was based on interpretability, flexibility and ability to handle multivariate data. After selecting an MBCFD method, we used it to identify trajectories (i.e., clusters) in the data considering the dimension of time. In order to determine

if the findings were clinically relevant, we graphically evaluated results in the context of previously found/proposed phenotypes within and outside of the CHECK cohorts with the help of clinical experts next to criteria of statistical fit and group size. Solutions using different sets of (multivariate) longitudinal outcomes combining different radiographic and clinical outcomes over time were explored and final solutions selected with the help of OA experts. The results of both methods were compared by using similarity metrics such as the Adjusted Rand Index (ARI) and statistical significance tests applied to baseline characteristics of the clusters such as gender, body mass index (BMI), age, and biochemical markers. Lastly, we present the extent of the contribution of the MBCFD method to detect clinically-relevant and statistically-significant phenotypes.

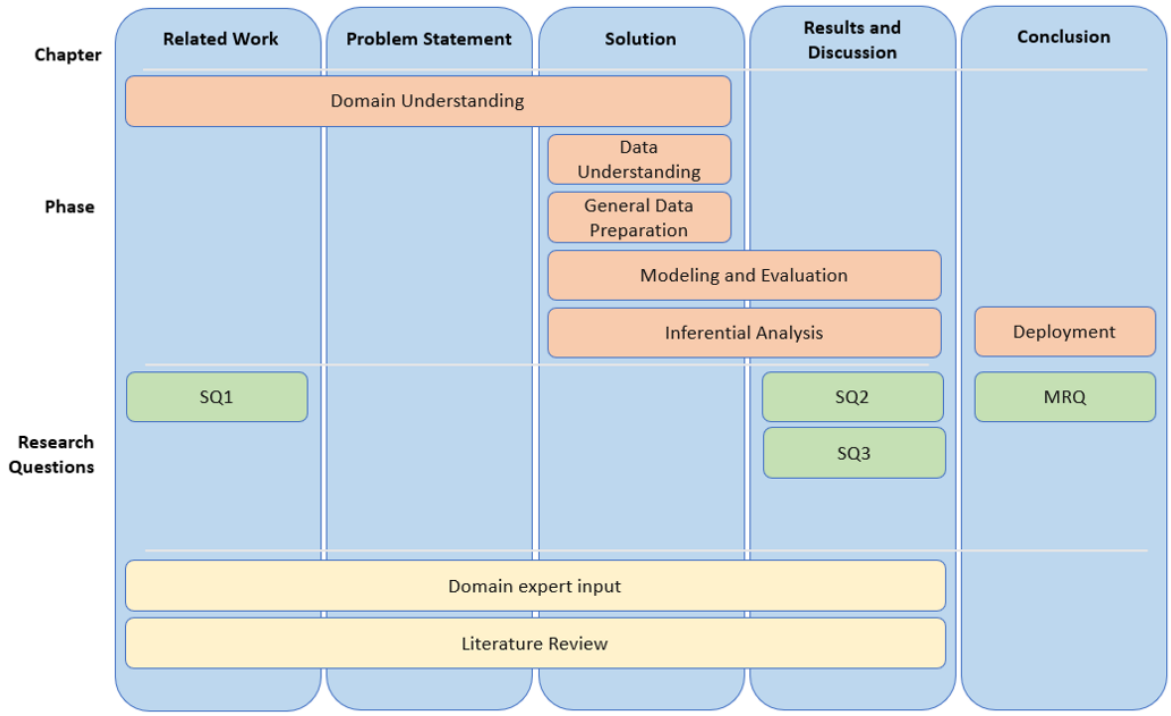


FIGURE 1: Overview of research approach.

Motivation

To the best of our knowledge, there has been no research conducted to date on whether any method that belongs to the FDA field of study can contribute to finding knee OA phenotypes. Previous work with CHECK data shows that patient phenotypes can be found. However, the work was based on a single or limited set of parameters and typically not the full 10-year follow-up. Therefore, the potential impact of these phenotypes on prognostication and treatment decisions is still limited. The majority of the studies used distance-based cluster analysis methods (see Appendix A). With our research, we intend to demonstrate that an MBCFD method can be useful to detect clinically-relevant and statistically-relevant phenotypes and to assess the performance vis-a-vis a clustering method widely used in the scientific literature.. In addition, by gaining a better understanding of OA heterogeneity (i.e., different phenotypes), we could potentially contribute to the design of clinical trials, prevention strategies, and treatments.

Research Aim

To define our research aim, we used the template defined by Wieringa (2014) because it is useful to "identify missing pieces of information . . . needed to bound your research problem" (p .16). The research aim is stated as follows:

- This research aims to *improve* OA phenotype understanding
- *by* testing an MBCFD method's ability to derive knee OA phenotypes and comparing to a clustering method's results to determine which method yields better performance
- *that satisfies* clinical relevance and statistical significance criteria
- *in order to* provide insights on the associations of OA parameters in individuals with symptoms of early-stage OA and potentially contribute to society by impacting trial design and the development of personalized prevention and treatment strategies for OA.

Research Questions

To address our research aim, the research is structured with a main research question (MRQ):

> **MRQ:** To what extent can model-based clustering for functional data (MBCFD) contribute to derive clinically-relevant and statistically-significant knee OA phenotypes?

Methods which belong to the MBCFD category can deal with data presented in the form of functions or curves and take into account the progression through time otherwise ignored by distance-based methods. By including the dimension of time and modeling the data as curves, we expect that an MBCFD method can contribute to derive the phenotype of an OA patient by yielding better performance than the most commonly-used method in scientific studies. We define performance by means of clinical relevance criteria and statistical significance between the derived groups' baseline characteristics. Clinically-relevant phenotypes can be distinguished based on differences between patients in the extent and course of their disease, the longitudinal associations between structural and/or clinical OA-related parameters, and markers of underlying etiologic mechanisms. Statistical significance can be determined by hypothesis testing to understand if our findings are unlikely to have happened by chance.

The **MRQ** is answered by the following four subquestions (SQs):

- **SQ1: Which are the most commonly-used methods in the literature being used to derive phenotypes from knee OA data?**

  We reviewed existing methods in the scientific literature to catalog the research conducted thus far. We computed the frequency to determine the top three methods. In addition, we created a list of the features used in these studies (**SQ1.1**) and which phenotypes were discovered. Lastly, we reviewed the literature to investigate which MBCFD methods are available for the analysis (**SQ1.2**).

  SQ1.1: Which are the characteristics used for identifying knee OA phenotypes?

  SQ1.2: Which are MBCFD methods that can be used for deriving knee OA phenotypes?

- **SQ2: How well does the selected MBCFD method perform at identifying clinically-relevant and statistically-significant knee OA phenotypes?**

  Based on the answers from **SQ1**, we performed the analysis with the selected MBCFD method and assessed the clinical relevance and statistical significance of the discovered subgroups. The input of domain experts was used to determine the clinical relevance of the subgroups throughout the interactive modeling and evaluation iterations' unstructured interviews. Statistical significance was explored with appropriate hypothesis testing.

- **SQ3: How does the MBCFD method perform compared to a non-functional clustering method?**

  We performed clustering analysis with the most common method discovered when answering **SQ1** and then compared the results with MBCFD clusters by assessing clinical relevance, statistical significance, and dissimilarity between clusters.

Research Methods

In order to approach our research problem, the CHECK (Wesseling et al., 2014) data set has been chosen. CHECK is a population-based cohort study of 1002 subjects with symptoms of knee and/or hip OA, with none to minimal radiographic signs at baseline. Over a ten year period, these individuals were followed for OA-related symptoms, physical, and radiographic signs. In addition, biochemical markers were obtained at baseline. CHECK has an especially low loss to follow-up given a targeted retention program (Wesseling et al., 2014).

We believe that an MBCFD method can be used to derive clinically relevant knee OA phenotypes and that this method outperforms HCA. To answer our research questions, we applied the CRISP-IDM method's six phases: domain understanding, data understanding, general data preparation, modeling and evaluation, inferential analysis, and deployment, which can be seen in Figure 1. These phases were implemented as follows:

1. **Domain Understanding**

   This phase spreads across chapters 1, 2 and 3 and consists of understanding the context, the problem at hand, project goals and requirements via the organization of topics and themes and conducting meetings with domain experts in context. In addition, we conduct the literature review to understand the state of the art. Meetings with domain experts are held to understand the specific problem and potential research gaps. Forthwith, potential research questions can be derived from the topics and themes which are required for choosing relevant data sources and the general data preparation phase. Under these circumstances, a list of topics arises which are subsequently categorized into the themes. Priority levels can be assigned to topics and/or themes. The motive behind the identification of topics is to discover information leading to potential analyses that can be conducted. Lastly, the outcome of this phase is a list of topics and themes.

2. **Data Understanding**

   This phase can be found in chapter 3 and consists of investigating the identified research themes from the previous phase. Understanding the data requires the selection of relevant data sources, and gaining lawful access to these data. Accordingly, data sources are summarized into data files that contain the source, name, type, structure, and number of records. The sources can stem from diverse internal and external repositories. Lastly, the outcomes of this phase are a description of the data and a data file table with a list of data sources/entities with types, structure, and number of records.

3. **General Data Preparation**

   This phase can be found in chapter 3 and refers to the preprocessing of the data to convert it into the appropriate format for exploratory analysis, and storing the data into a database. The most relevant steps in preparing the data are organizing, transforming, cleaning, integrating, sampling, reducing and/or discretizing the data. The outcome of this phase is clean and understandable data sets ready to be consumed by the next phase.

4. **Modeling and Evaluation**

   This phase can be found in chapters 3 and 4. It is interactive in nature since several iterations have to be performed when, with the aid of data visualization software, the data is modeled and presented to medical experts who provide their feedback to be used for the subsequent iteration. These iterations or cycles can widely vary in number. With the guide of experts in the field of OA and their direct feedback, we aim to reach a consensus on the clinically-relevant phenotypes derived with the MBCFD and traditional clustering methods. The outcomes of this phase are the selected and prepared database, visualizations, and feedback from domain experts.

5. **Inferential Analysis**

   This phase can be found in chapters 3 and 4 and refers to testing hypotheses and statistical significance, making inferences about the data. We aim to discover the statistically-significant characteristics of the derived phenotypes by comparing the baseline characteristics of each cluster such as BMI, age, and biochemical markers.

6. **Deployment**

   During this phase, the implementation of results that were obtained and confirmed occur. After debating the results with involved domain experts, a consensus is reached, which increases the prospect of success. The deployment phase focuses on the implementation of results, which in our case will take the form of a scientific article, a process-deliverable diagram (PDD), and this thesis report. The PDD is created to detail the phases, activities, and deliverables (concepts) of our solution and guide researchers through similar projects in the future.

CRISP-IDM is complemented by the consensus-based framework for conducting and reporting OA phenotype research developed by Van Spil et al. (2020). This framework entails reporting recommendations with regards to general study characteristics, study population, data collection, statistical analysis, and appraisal. We focused on some of the recommendations regarding statistical analysis. The research is being performed at the Rheumatology and Clinical Immunology Department of University Medical Center Utrecht (UMCU) in Utrecht, The Netherlands.

We began by investigating machine learning methods commonly used in the literature for knee OA phenotypes. Then, we defined the characteristics used in the respective analyses. Moreover, we investigated potential MBCFD methods and selected funHDDC. By following the CRISP-IDM method, we fulfilled six phases which included 30 iterations of data exploration in which we determined that funHDDC can detect clinically-relevant and statistically-significant knee OA phenotypes for the univariate case. However, for the multivariate case, results were limited to clinical relevance. Lastly, by comparing to results from HCA clusters, we determined that funHDDC outperforms HCA in the univariate case but not in the multivariate case.

Our contributions can be summarized as follows:

- we used early-stage OA data - which could be better suited to conduct phenotype research as opposed to established or end-stage OA (Whittle et al., 2016)
- we used longitudinal (10-year) data versus shorter follow-up times
- we used parameters from different domains such as clinical, imaging, demographic and biochemical markers data; which may be useful to define phenotypes (Deveza, Nelson, and Loeser, 2019)
- we explored the performance of MBCFD for deriving knee OA phenotypes and compared to a commonly-used method in literature studies
- the research was performed with the assistance of researchers who are familiar with the CHECK cohort, OA etiology and clinical practice, and required analytical methods

Thesis Outline

The remainder of this thesis is structured as follows. First, to provide necessary background knowledge and to put the problem in context, chapter 1 presents the related work. Next, chapter 2 describes the precise issue the research addresses in more detail, shows the relevance of the problem and highlights the research gap. Subsequently, chapter 3, presents the solution, chapter 4, the results and discussion, and chapter 5, the conclusion. Appendix A showcases studies that investigated knee OA phenotypes. Appendix B details the steps taken during the data exploration steps of CRISP-IDM. Appendix C presents the baseline characteristics tables with statistical significance results. Appendix D contains clustering evaluation plots and tables. Lastly, appendix E presents examples of the code used in the project. The thesis outline is represented in Figure 2.

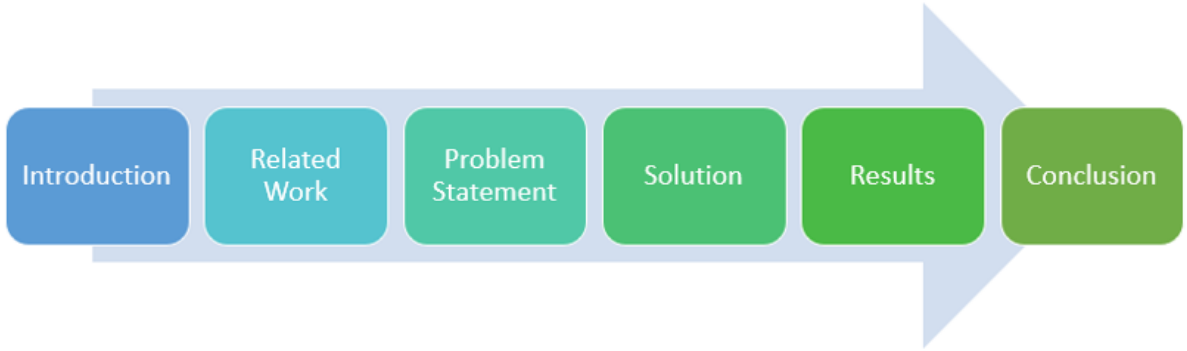


FIGURE 2: Overview of thesis outline.

Chapter 1

Related Work

Recently, the concept of osteoarthritis (OA) heterogeneity has been gaining acceptance as etiological mechanisms vary between OA patients causing differences in symptomatology and disease progression. There is significant variation in disease prognosis between patients, with some patients enduring progression while others remain stable (Karsdal et al., 2015). Moreover, it is plausible that heterogeneity is most perceptible in early-stage OA. Thus, the highest opportunity to provide personalized medicine is most likely early in the disease course. Moreover, at this stage, symptoms are less severe and no significant joint damage has yet occurred, thus treatment might be more effective. Additionally, etiological mechanisms might be explicitly different between patients early in the disease course and less so later, as progressively more factors converge over time. These phenomena may have an inconsistent impact on patients' reactions to treatments, and potentially provide clarification for the lack of success of OA clinical trials (Bruyere et al., 2015). Better understanding of the entire range of factors that are implicated in OA heterogeneity is essential to advance the consolidation of knee OA phenotypes (Deveza et al., 2017), which are the observable (i.e., distinguishable by inspection) characteristics of an individual (Johannsen, 1911) resulting from a combination of environmental and genetic factors (Deveza, Nelson, and Loeser, 2019).

The literature evidences that few studies have attempted to define a multidimensional stratification of phenotypes, and there is myriad research opportunities in testing the prospective validity of the subgroups using longitudinal outcomes (Deveza et al., 2017). Appendix A summarizes prospective studies investigating potential knee OA phenotypes based on the systematic literature review by Deveza et al. (2017) as well as an updated list of studies from 2018 to 2020.

Current and future research need to continue to explore potential phenotyping methods as these may be the answer to improving palliative treatment and creating interventions that improve the quality of life for homogeneous subgroups considering the heterogeneous nature of OA. The importance of continuing to explore existing and not-yet-discovered phenotypes cannot be stressed enough as this could change the course of OA health care by guiding personalized, life-changing, and disease-modifying interventions.

Identifying potential relevant literature was performed by searching particular keywords on Google Scholar and PubMed. An initial set of articles is selected by reviewing the title and abstracts of the first 100 results. The 100-article set was subsequently reviewed in depth which led to the exclusion of tangential material. Eventually, through the use of backward and forward *snowballing* (Wohlin, 2014), additional materials were identified.

1.1 Osteoarthritis

OA is defined as:

> **Definition** (*Osteoarthritis*).
> The most common chronic joint disease (Bijlsma, Berenbaum, and Lafeber, 2011), characterized by pain, functional disability, and limited quality of life (Arden and Nevitt, 2006).

OA can be characterized by joint symptoms, by structural pathology such as evidenced on X-rays or by their combination. The prevalence of OA is increasing due to the aging population and widespread obesity (Bijlsma, Berenbaum, and Lafeber, 2011). Moreover, OA is a progressive disease that can affect all joint structures and lead to joint failure by impacting articular cartilage, subchondral bone, synovium, meniscus, muscle, capsule, and ligaments (Bruyere et al., 2015). OA may occur due to a broad spectrum of factors such as trauma, hereditability, and biomechanical and metabolic issues; multiple mechanisms can play a part in the perception of pain (Castañeda et al., 2014). In general, OA develops progressively over several years, however, symptoms can remain stable for long periods. Figure 1.1 shows an illustration comparing a healthy knee with a knee affected by OA. The pathogenesis of OA is regarded to be a combination of factors and to be different between patients, although the mechanisms of genesis and progression remain unidentified, some predisposing risk factors have been identified for both knee and hip OA. For the occurrence of the disease, these are common risk factors: age, physical activity, body mass index, obesity, previous injury, intense sport activities, and genetics. On the other hand, for the progression of the disease, age, and intense sport activities are common risk factors. However, pain and loss of function are the main clinical features that lead to treatment (Bijlsma, Berenbaum, and Lafeber, 2011).

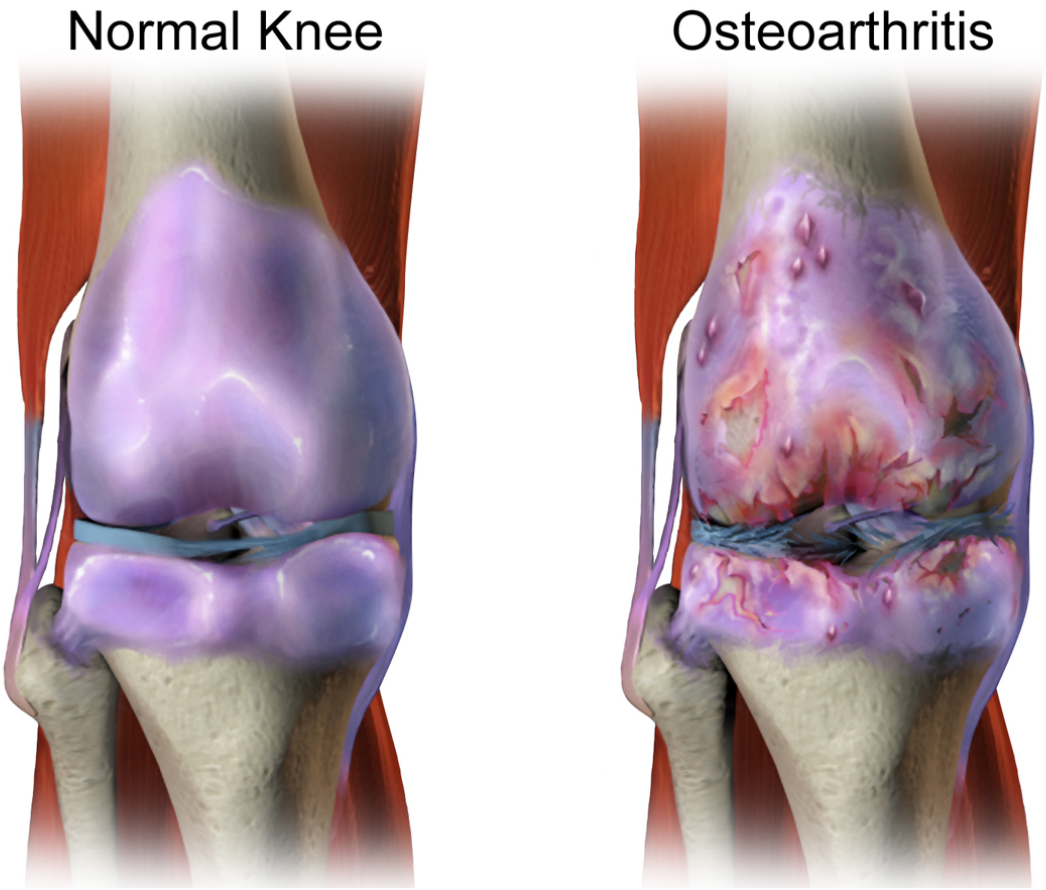


FIGURE 1.1: Illustration of knee osteoarthritis. From Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine (2): 10.

Knee OA can make it difficult to do many daily common activities, such as walking or exercising, as the cartilage in the knee can progressively wear away causing the space between the bones to diminish in size. In a clinical setting, the diagnosis of OA is usually based on clinical complaints and subsequently confirmed by radiographic evaluation of structural damage

(Bedson and Croft, 2008). There are three treatment modalities: non-pharmacological, pharmacological, and surgical. In many cases, these modalities are used in combination to tailor treatment to specific scenarios. Furthermore, since pain is multi-factorial, it is also affected by comorbidities, such as sleeping problems, loneliness, and mood disorders; therefore, improvement of mental and emotional wellbeing is of utmost importance (Geenen and Bijlsma, 2010).

1.1.1 Early Symptomatic OA

For prevention and early intervention purposes, it is crucial to diagnose OA at an early stage and identify its prognostic signs (Wesseling et al., 2014). However, OA is typically diagnosed at a late stage when structural damage is considered irreversible and, therefore, treatment focuses mostly on relieving pain (White et al., 2010). To understand the disease process and for the development of satisfactory disease-modifying treatment alternatives for OA patients, a greater emphasis has to be made on identifying potentially high risk individuals so measures can be taken to prevent irreversible damage. In early OA, pain and stiffness could mask other symptoms, treatment should therefore focus on decreasing pain and stiffness and on the maintenance and betterment of functional abilities. Correspondingly, prevention of progression of joint damage and improvement of quality of life should be the aim (Bijlsma, Berenbaum, and Lafeber, 2011). By identifying clinically-relevant phenotypes, trial design and the development of tailored prevention and targeted treatment strategies can be positively impacted in order to identify more efficacious solutions (Deveza et al., 2017).

1.1.2 OA Phenotypes

OA phenotypes are defined as:

> **Definition** (*Osteoarthritis Phenotypes*).
> "Subtypes of OA that share distinct underlying pathobiological and pain mechanisms and their structural and functional consequences" (Van Spil et al., 2020, p. 4).

Discovering knee OA phenotypes requires employing clustering techniques to identify clusters of individuals with similar characteristics with the long-term goal of personalizing patient management (Pinto et al., 2015). Deveza et al. (2017) performed a systematic literature review to identify which features are important for phenotyping knee OA and found that clinical phenotypes are investigated more frequently, followed by laboratory, imaging and etiologic phenotypes. Additionally, the authors found eight studies that defined subgroups based on outcome trajectories (pain, function and radiographic progression trajectories) with only five studies including characteristics from multiple domains. Evidence was found to suggest that pain sensitization, psychological distress, radiographic severity, body mass index (BMI), muscle strength, inflammation, and comorbidities are related to clinically distinct phenotypes. Gender, obesity and other metabolic irregularities, pattern of cartilage damage, and inflammation may be involved in describing distinct structural phenotypes. A handful of studies researched the phenotypes' external validity or their potential validity using longitudinal outcomes.

Previous studies have grouped OA patients into phenotypes from diverse perspectives by using different sets of characteristics to determine phenotypes, such as experimental pain sensitivity (Cardoso et al., 2016), imaging (Roze et al., 2016), biochemical markers (Zhang et

al., 2014; Van Spil et al., 2012), comorbidities (Murphy et al., 2011) and clinical characteristics (Knoop et al., 2011). These studies used trajectories of clinical or structural progression (outcome-based definitions) or baseline characteristics with the subsequent association of the phenotypes with outcomes.

Multiple methods have been employed to identify OA phenotypes, such as logistic regression (Nelson et al., 2013), HCA (Iijima et al., 2015; Cardoso et al., 2016), expert opinion (Castañeda et al., 2014), k-means clustering (Knoop et al., 2011; Elbaz et al., 2014), latent class analysis (Kittelson, Stevens-Lapsley, and Schmiege, 2016; Waarsing, Bierma-Zeinstra, and Weinans, 2015), and principal component analysis (Meulenbelt et al., 2007; Heard et al., 2013). However, HCA is the most common method besides pre-defined phenotypes. Appendix A provides a summary of the findings from Deveza et al. (2017) by means of an overview of the phenotype research performed in prospective studies investigating knee OA characteristics and outcomes as well as the baseline variables that more frequently predicted worse trajectory outcomes. Baseline variables that seem to predict worse outcomes included high BMI, lower education, more severe symptoms and radiographic disease at baseline, psychological factors, and presence of other comorbidities including accompanying hip pain. Appendix A presents characteristics, authors, features, methods, and phenotypes discovered. There are three main categories observed in grouping the data: clinical, imaging (radiography), and laboratory. For clinical data, the methods used for discovering phenotypes are hierarchical clustering, k-means clustering, latent class analysis, and expert opinion. The characteristics used for grouping the data relate to pain, knee joint alignment, metabolic profile, comorbidities, gait parameters, psychological profiles, and mechanistic factors. The phenotypes found range from two to five in number. More details can be seen in Table A.1. For imaging data, the methods used for finding subgroups are expert opinion and latent class analysis. The characteristics explored relate to imaging features, knee chondrocalcinosis, knee joint compartment, and MRI-detected denuded bone areas. The phenotypes found range from two to four in number. More details can be seen in Table A.2. Lastly, for laboratory data, the methods used are expert opinion, principal component analysis, and k-means. The characteristics relate to biochemical market patterns, synovial fluid, inflammation, and markers of bone and cartilage metabolism. More details can be seen in Table A.3.

To complement the Deveza et al. (2017) study, we performed a semi-systematic literature review (Snyder, 2019) to close the knowledge gap and discover what additional research has been conducted from January 2018 to March 2020 regarding the use of machine learning for finding knee OA phenotypes. We identified relevant articles and abstracts in a search of PubMed and Google Scholar for English language journal articles. An additional resource screened was the CHECK research website [1]. The search strategy for PubMed can be seen in Table 1.1. The search terms were based on the keywords used by Deveza et al. (2017).

TABLE 1.1: Search Strategy (PubMed).

#	**Search strategy**	**Results**
1	(OA/) OR (OA, knee/)	26,277
2	arthrosis	33,682
3	(osteoarthr*) OR (osteo-arthr*)	15,603
4	(degenerative) AND (arthritis)	866
5	phenotype*	83,494
6	(subgroup*) OR (sub-group*)	47,199

[1] https://www.check-onderzoek.nl/publication-presentation/scientific-publications/

Table 1.1 continued from previous page

#	Search strategy	Results
7	(subtype*) OR (sub-type*)	33,928
8	(subset* OR sub-set*)	34,669
9	(subpopulation*) OR (sub-population*)	9,141
10	cluster*	74,749
11	Phenotype/	107,388
12	(knee/) OR (OA, knee/)	26,277
13	1 or 2 or 3 or 4	41,801
14	12 and 13	9,959
15	5 or 6 or 7 or 8 or 9 or 10 or 11	281,113
16	14 and 15	744
17	limit 16 to humans	448
18	limit 17 to "review" articles	30
19	17 not 18	418

Overall, 418 and 961 records were found on PubMed and Google Scholar, respectively. Studies were eligible if they:

- Have the goal of identifying knee OA phenotypes
- Use a machine learning algorithm

The titles and summaries of the first 250 records sorted by best match from PubMed were screened. Similarly, the titles and summaries of the first 250 results from Google Scholar were screened. After the initial screening of titles and summaries, 167 abstracts were reviewed and 52 studies were selected. Subsequently, after assessing articles for eligibility, 29 articles were excluded for not meeting the eligibility criteria. For the set of 20 studies that met the selection criteria, the methods used were cluster analysis, regression analysis, DWD, latent class analysis, principal component analysis, hierarchical cluster analysis, latent class growth analysis, group-based trajectory modeling, and support vector machine. The characteristics relate to pain, clinical and radiographic measures, biochemical markers, synovial fluid, quality of life, depression, biomechanical measures, comorbidities, and gene expression. The number of phenotypes found ranged from two to six. Table A.4 in Appendix A shows the characteristics, author, features, methods, and phenotypes of the 20 articles that met the selection criteria. The process and results from the semi-systematic literature review can be seen in the PRISMA (Moher et al., 2009) flowchart in Figure 1.2.

1.2 Answers to Research Questions SQ1 and SQ1.1

Our first research subquestion poses the following:

1.2.1 SQ1: Which are the most commonly-used methods in the literature being used to derive phenotypes from knee OA data?

To answer SQ1, we extracted the most commonly-used methods used for deriving knee OA phenotypes from the scientific literature. We computed the frequency of the methods based on a total of 46 studies. The top methods are:

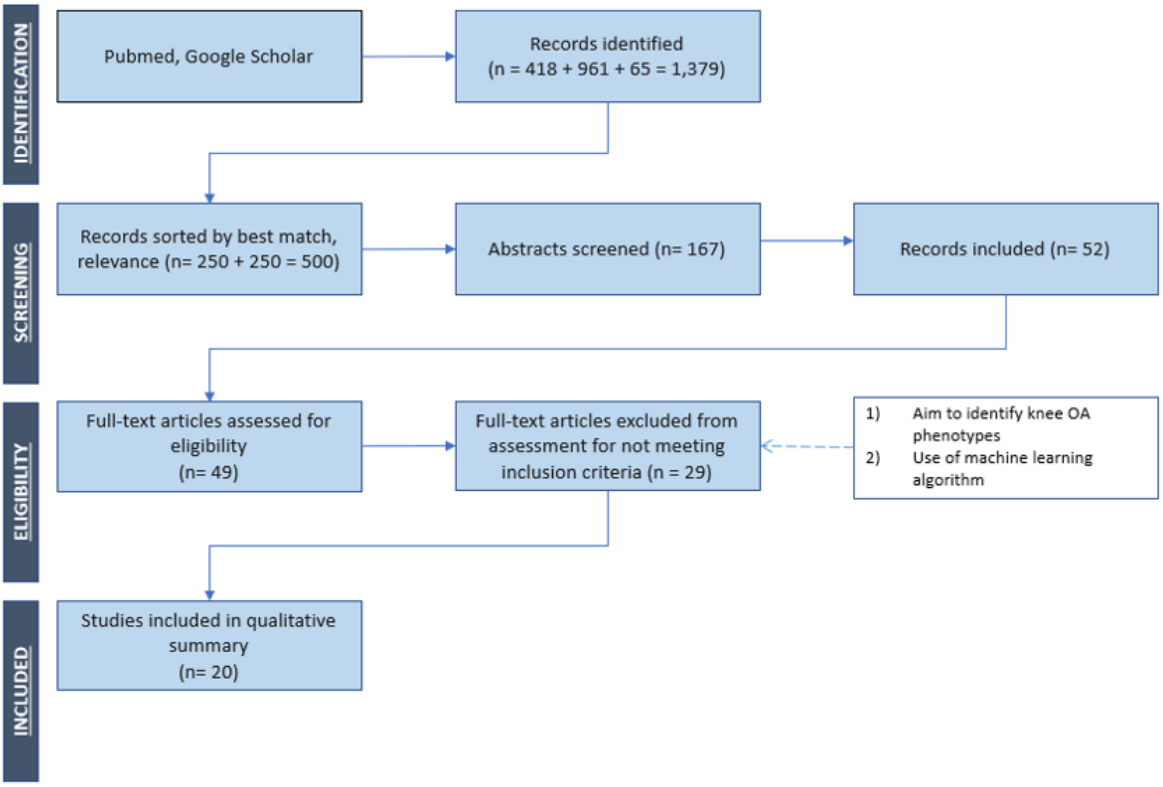


FIGURE 1.2: PRISMA flowchart describing the literature review of OA phenotype methods from 2018 to 2020.

1. Hierarchical cluster analysis (HCA) (24%). HCA, as the name states, seeks to build a *hierarchy* of clusters based on similarity within clusters and dissimilarity between clusters. HCA typically uses a distance matrix to determine the groups. More details about this method can be found in section 1.3.1.

2. Latent class analysis (LCA) (15%). LCA is a model-based clustering approach that identifies unobserved classes by grouping multivariate data into *latent classes* recognizing hidden patterns (using conditional probability) that associate the observations (Vermunt and Magidson, 2002).

3. k-means (9%) and logistic regression (9%). k-means is a hard clustering approach since each observation can belong only to one cluster or partition. Logistic regression uses a logistic function to model discrete (e.g., 0/1) data.

All commonly used methods belong to the unsupervised learning category, with the exception of logistic regression which is a supervised classification method. HCA has some advantages over k-means because it does not need a priori specification of the number of clusters and provides a graphical, tree-based representation of the groupings, called a *dendrogram*.

SQ1 has two further subquestions: SQ1.1 and SQ1.2. We answer SQ1.2 in section 1.3.3.

1.2.2 SQ1.1: Which are the characteristics used for identifying knee OA phenotypes?

From Deveza et al. (2017) and our semi-systematic literature review, we can list the following characteristics:

- pain
- psychological profiles
- comorbid-symptoms profile
- clinical characteristics
- knee joint alignment
- metabolic profile
- gait parameters
- mechanistic factors
- knee chondrocalcinosis (i.e., calcium pyrophosphate build up in the joints)
- MRI-detected denuded bone areas
- imaging
- knee joint compartment
- biochemical markers
- inflammatory profile
- synovial fluid profile
- gene expression
- quality of life
- depression
- functional capacity
- comorbidities

Therefore, the most frequently-used characteristics in OA phenotype research are **pain, imaging measures (X-ray), clinical measures, biochemical markers, and gene expression.**

In addition, Dell'Isola et al. (2016) found, through qualitative analysis, six main sets of variables that suggest the existence of six phenotypes:

> (i) chronic pain in which central mechanisms (e.g. central sensitisation) are prominent; (ii) inflammatory (high levels of inflammatory biomarkers); (iii) metabolic syndrome (high prevalence of obesity, diabetes and other metabolic disturbances); (iv) bone and cartilage metabolism (alteration in local tissue metabolism); (v) mechanical overload characterised primarily by varus malalignment and medial compartment disease; and (vi) minimal joint disease characterised as minor clinical symptoms with slow progression over time. (p. 1)

Two of the closest works to our research come from Deveza et al. (2019) and Nelson et al. (2019). On the one hand, Deveza et al. (2019) analyzed 2-year data (n = 1,014) from the Osteoarthritis Initiative (OAI) and applied latent class growth analysis (LCGA) to identify trajectories using demographic, clinical and radiographic data. LCGA is a model-based clustering method useful for finding "groupings of individuals who share similar longitudinal data patterns to determine the extent to which these patterns may relate to variables of interest" (Berlin, Parra, and Williams, 2014, p. 3) which makes this method appropriate for discovering phenotypes by finding homogeneous subpopulations within the overall heterogeneous population. However, as is the case with any statistical method that uses discrete data as input, we must be very cautious when discretizing continuous measures to reasonably ensure the resulting values do not imply a significant loss of information. This limitation could be overcome when using an FDA method since the information between time points remains a part of the ecosystem. Perhaps Latent Profile Analysis is better suited for phenotype research since it can handle continuous data. Another limitation of Deveza et al. (2019) could be the use of 2-year data for a slowly progressive disease. The question remains whether two years' worth of data is enough for accurately representing progression and inferring phenotypes. On the other hand, Nelson et al.

(2019) used 10-year demographic, imaging, and biochemical data (n = 597) from the Foundation for the National Institutes of Health (FNIH) Osteoarthritis (OA) Biomarkers Consortium. Nelson et al. (2019) applied Distance-Weighted Discrimination, a supervised distance-based learning method that allows maximal separation of data points by class and treats each vector of features as a single data object. In contrast to this approach, we are applying an MBCFD method without having any prior knowledge on the group labels and using early-stage OA data instead of established OA data. Additionally, we are transforming each variable's discrete data points into functions, which could allow us to see complex trends not detected by traditional distance-based clustering algorithms.

1.3 Unsupervised Machine Learning

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) which relies on experience-based computational methods to optimize task performance or derive accurate predictions (Mohri, Rostamizadeh, and Talwalkar, 2018). A typical example of ML is predicting the class of an unseen observation based on a set of pre-labeled random observations. However, not all data are labeled in real-world applications and this is where unsupervised learning has its role.

Unsupervised learning is defined as:

> **Definition** (*Unsupervised Learning*).
> Unsupervised learning is a set of statistical tools suitable for a scenario where we only have a set of features $X_1, X_2, ..., X_p$ measured on n observations but we are not aiming for prediction due to the absence of response variable Y. "The goal is to discover interesting things about the measurements on $X_1, X_2, ..., X_p$." (James et al., 2013, p. 373)

Unsupervised learning is a subfield of machine learning that focuses on processing data that have not been classified with a *ground truth label*. The goal of unsupervised learning is to model the implicit structure of the data to learn more about it without receiving feedback. Unsupervised learning algorithms can have many applications in different fields. Examples of unsupervised learning methods are clustering data points via similarity metrics and dimension reduction or feature selection. Two of the most commonly used unsupervised machine learning methods are principal component analysis (PCA) and cluster analysis. Regardless of the method chosen to conduct the analysis, the biggest challenge in unsupervised machine learning can be to assess whether results are relevant to the domain since we do not know the true answer, thus it is recommended to pair these analyses with subject-area expertise to contribute to the validity of the findings (Deveza, Nelson, and Loeser, 2019). Moreover, it is also challenging to assess the results since there are no widely accepted validation tools (James et al., 2013).

1.3.1 Cluster Analysis

Cluster analysis is an unsupervised machine learning method which is defined as:

Definition (*Cluster Analysis*).
An unsupervised machine learning method with the goal of "finding meaningful groups in the data. The purpose (of cluster analysis) is to find groups whose members have something in common that they do not share with members of other groups" (Bouveyron et al., 2019, p. 1).

Clustering is useful for exploratory data analysis as it identifies patterns in unlabeled data by systematically grouping objects (Aghabozorgi, Shirkhorshidi, and Wah, 2015). Examples of systematic grouping of objects date back to the 1700s with the taxonomy of Linneaus where he classified and labeled organisms. In the 1990s, there was an increased interest in clustering emerging data such as genetic microarray data, barcode data, websites, medical images, among many others (Bouveyron et al., 2019). For the history of cluster analysis before the 1990s, see Blashfield and Aldenderfer (1988).

The majority of the prior clustering methods were algorithmic and heuristic in nature, finding latent groups in the data by extracting measures of similarity between objects from their observed attributes, e.g., age, height, gender (Bouveyron et al., 2019). As definitions of similarity vary from one clustering mechanism to another, usually the concept of similarity relies on *distance*. There are several well-known types of distance, such as the Euclidean or Manhattan distances, but the underlying idea is that the data points should end up in clusters that are dissimilar between and similar within. The discovery of clusters in data sets using pattern similarity is extensively relevant when unearthing *actionable insights*. An example of an application is when using clustering in DNA micro-array analysis by means of expressing patterns of thousands of genes. These data are arranged in matrix form where each row is a different gene, columns represent samples (i.e., tissue) and the values in the cells describe an observed data point. Since particular genes contribute to specific diseases, researchers aim to find which genes are expressed, thereby providing a phenotype for a sample. This application's basis is grouping observations which have consistent behaviors (Wang et al., 2002). A challenge to overcome with clustering techniques is finding new similarity models to group longitudinal data since well-known distance measures may not fully capture the relationships among the objects with the caveat that sequence integrity is ignored in the analysis.

Clustering has been studied thoroughly for decades as cluster analysis was first introduced by Driver and Kroeber (1932) in the anthropological sciences. The latest clustering and classification research has focused on Bayesian regularization methods, non-Gaussian model-based clustering, cluster merging, variable selection, semi-supervised classification, robust classification, clustering of functional data, text, and images, and co-clustering (Bouveyron et al., 2019). More recently, cluster analysis has evolved to encompass model-based clustering given the diversity of the data, new scientific questions, and very large data sets.

Hierarchical Cluster Analysis

Hierarchical cluster analysis (HCA) was developed to overcome some of the disadvantages that accompany partitional-based clustering methods such as k-means as they usually require a priori definition of K number of clusters. HCA was designed with a more flexible approach in mind for clustering the data points. HCA can be categorized into agglomerative and divisive methods. The agglomerative approach starts by creating singleton clusters with only one data point per cluster and continues adding two clusters at a time to create a bottom-up hierarchy of clusters. In contrast, the divisive approach starts with one all-encompassing cluster and divides it continuously into two groups creating a top-down hierarchy of clusters (Aggarwal, 2014). HCA results

are presented in the form of a tree-like diagram which represents the distance-based associations between data points. Figure 1.3 shows an example of applying agglomerative hierarchical clustering to a sample dataset (created by the author) with 20 patients and their yearly BMI over 10 years. The resulting dendrogram shows the hierarchical relationship between the patients. Each *leaf* of the dendrogram represents one of the 20 patients. When we move up the tree, the leaves start joining together to form *branches*, the lower these joins happen, the more similar the groups of patients are to each other. Observations that join later (up the tree) are the most different. The vertical axis represents the height of the joins and indicates the similarity between two observations. In other words, we read the similarity between two observations based on the exact location on the vertical axis where branches containing those two observations are joined first (James et al., 2013).

A challenge to overcome with HCA methods is that Gaussian mixture models typically result in clusters with convex geometric shape (e.g., when using Euclidean distance) which might not be ideal when attempting to cluster trajectories or non-convex shapes. The construct of similarity between a pair of observations can be extended by developing the concept of *linkage*. The four most common types of linkage (complete, average, single, centroid) and their definitions can be found in Table 1.2. Ward distance (Ward, 1963), which computes the minimum within-cluster variance, is also available.

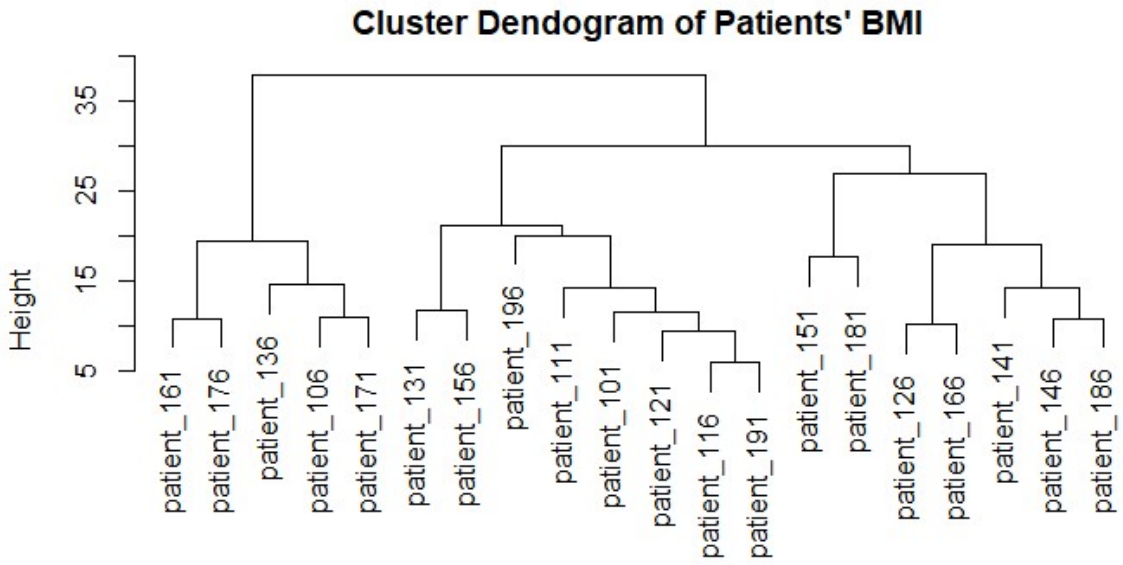


FIGURE 1.3: Example of a cluster dendrogram.

TABLE 1.2: The four most commonly used types of linkage in hierarchical clustering, adapted from James et al., 2013, p. 395.

Linkage	**Description**
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.

Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

Clustering Evaluation

Unsupervised clustering can be difficult to evaluate due to the lack of ground truth labels. If ground truth labels are available, then supervised evaluation methods can be used (James et al., 2013). For example, one can compile a confusion matrix and calculate metrics such as accuracy, sensitivity, or specificity (Aggarwal, 2014). Conversely, despite the lack of ground truth labels in unsupervised problems, some factors that can be evaluated are clustering tendency, number of K optimal clusters, and clustering validity (Han, Pei, and Kamber, 2011). Moreover, intercluster statistical significance tests can performed.

Statistical significance. In the field of inferential statistics, hypothesis testing provides methods to extract information from a representative sample to draw conclusions about a population. Hypothesis testing requires a null hypothesis (i.e., stating there is no effect or that the effect was due to chance) and an alternative hypothesis (i.e., there *is* an effect). We test statistical significance to understand if our findings are unlikely to have happened by chance. When the null hypothesis is rejected, we can state we have discovered significant results. When we fail to reject the null hypothesis, we do not achieve statistical significance. Typically, hypothesis testing is performed by setting a threshold for the p-value, known as α or *level of significance*, and decide there is statistical significance by rejecting the null hypothesis when p-value $\leq \alpha$. Usually α is set at 0.05 (Black, 2019).

There are several hypothesis tests that fit into two main categories: parametric and non-parametric tests. Since we do not wish to make any assumptions about the population, we focus the scope of this study on non-parametric tests. When the objective is to understand the difference between the means of the clusters, analysis of variance is appropriate. The choice of test will also depend on the type of data. One-way ANOVA, developed by the statistician Ronald Fisher, is a parametric test and its non-parametric equivalent is the Kruskal-Wallis rank sum test which was developed in 1952 by W. Kruskal and W. Wallis. The Kruskal-Wallis rank sum test is used to investigate whether three or more samples originate from the same populations with no assumption about population distribution. Moreover, the rank-based Kruskal-Wallis rank sum test assumes group independence and random selection of observations. The hypotheses according to the Kruskal-Wallis rank sum test are:

H_0 : the populations are identical.

H_1 : at least one of the populations is different.

The equation used to compute the K statistic is:

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^{c} \frac{T_j^2}{n_j} \right) - 3(n+1),$$

where c is the number of groups, n is the total number of items, T_j is the total of ranks in a group, n_j is the number of items in a group, and with $K \approx \chi^2$ with df $= c - 1$.

Clustering tendency attempts to ascertain whether the data contains uniformly distributed points (random structure), otherwise the clusters identified may not be meaningful. In other words, the clustering tendency represents the *clusterability* of the data at hand. To address this problem, a statistical test for spatial randomness of a variable called *Hopkins statistic* (Hopkins and Skellam, 1954) can be used to measure the probability of the data being generated by a uniform data distribution. The Hopkins statistic is calculated as:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}, \tag{1.1}$$

where n is the number of sample points from the distribution, y is the distance between each data point and uniformly randomly distributed data points, and x is the distance between a randomly chose data point and its nearest neighbor in N. If N were uniformly distributed, then the Hopkins statistic equation denominator elements would be similar, making H approximately 0.5. On the contrary, if N were highly skewed then $\sum_{i=1}^{n} y_i$ would be considerably smaller than $\sum_{i=1}^{n} x_{,i}$, making H approximately 0 (Han, Pei, and Kamber, 2011). Consequently, if $H < 0.5$, it is improbable that N has statistically significant subgroups (Tan, Steinbach, and Karpatne, 2019).

Number of K optimal clusters is quite an important parameter in the analysis since missing the optimal K could mean loss of granularity by obscuring subgroups if K is too low or representing each data point as a cluster if K is too high. There are two well-known approaches to finding the optimal K, one is via domain knowledge and the other is a data-driven approach. Domain knowledge might be enough when the prior expert knowledge clearly gravitates towards a specific number of clusters. However, this is not always the case. If domain knowledge is not sufficient, then there are distance-based options to finding the optimal K which focus on compactness (how close are the objects within a cluster) and separation (how apart is a cluster from the others). There is a myriad of distance-based metrics used to measure compactness and separation, such as the elbow method, the average silhouette method, and the gap statistic. The elbow method focuses on within-cluster variance whereas the average silhouette method Fox (1991) is formulated by considering intra- and inter-cluster distances. The gap statistic "uses the output of any clustering algorithm (e.g. k-means or hierarchical), comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution" (Tibshirani, Walther, and Hastie, 2001, p. 1).

From the perspective of a model-selection problem, traditional approaches to select the number of clusters include the Akaike information criterion (Akaike, 1974), the Bayesian information criterion (Schwarz, 1978), and the integrated completed likelihood criterion (ICL) (Biernacki, Celeux, and Govaert, 2000). However, for mixture models, the most common criterion is BIC which is proportional to AIC, but it tends to penalize complex models more heavily, gravitating towards selecting simpler models. The BIC can be computed with the maximum log-likelihood value, the number of model parameters, and the number of observations, which

the criterion uses to penalize the log-likelihood via model complexity. Then, the model maximizing the criterion is chosen. The following equation computes the BIC for fitted model objects for which a log-likelihood value can be obtained:

$$BIC = l\left(\hat{\theta}\right) - \frac{m}{2} \times \log(n), \tag{1.2}$$

where $l\left(\hat{\theta}\right)$ is the maximum log-likelihood, m is the number of model parameters, and n is the number of observations (Schmutz et al., 2020). Maximum likelihood estimation finds the parameters that best fit the data by maximizing the probability of observing that data. Thus, "the most reasonable values for θ are those for which the probability of the observed sample is largest" (Hastie, Tibshirani, and Friedman, 2009, p. 31).

Clustering quality refers to how well the clustering exercises have performed and can be characterized by diverse measures. For distance-based clustering, the optimal groups have minimal intra-cluster distance and maximal inter-cluster distance. There are two main approaches to define clustering quality: *extrinsic measures* which require ground truth labels and *intrinsic measures* do not require ground truth labels. Extrinsic methods qualify clustering quality by satisfying the following criteria: (i) cluster homogeneity, (ii) cluster completeness, (iii) rag bag, and (iv) small cluster preservation. Extrinsic measures are out of the scope of our work because they require ground truth labels. Intrinsic measures evaluate clustering results by assessing the separation and compactness of clusters (Han, Pei, and Kamber, 2011). A similarity metric for objects in the data is the silhouette coefficient (Kaufman and Rousseeuw, 2009), which measures cohesion (within-cluster distance) and separation (between-cluster distance). This coefficient can be computed for any distance-based clustering exercise. The silhouette coefficient of $n \in N$ is defined as:

$$s(\mathbf{n}) = \frac{b(\mathbf{n}) - a(\mathbf{n})}{\max[a(\mathbf{n}), b(\mathbf{n})]}. \tag{1.3}$$

The resulting value ranges from -1 to 1. The value of $a(\mathbf{n})$ represents the cohesion of the cluster where smaller values mean more compact clusters. Moreover, $b(\mathbf{n})$ represents the degree of separation, where largest values mean more separation. Hence, when $s(\mathbf{n})$ reaches 1, it means that the cluster which contains $\mathbf{n}$ is compact and separated from other clusters, which is usually the goal. In contrast, when $s(\mathbf{n})$ is negative, it can be interpreted as n being closer to objects in other clusters than same-cluster members (Han, Pei, and Kamber, 2011).

Another measure used to assess agreement between two clustering exercises is the adjusted Rand index (ARI). The Rand index is defined as:

$$Rand = \frac{a+d}{T}, \tag{1.4}$$

where a is the number of pairs of observations belonging to the same cluster, and b is the number of pairs of observations not belonging to the same group in both clustering exercises. T is the total number of pairs. The values of the Rand index range between 0 and 1, with 1 meaning perfect agreement. However, the Rand index largely depends on the number of clusters in the two clustering exercises being compared. To overcome this challenge, the ARI was proposed by Hubert and Arabie (1985) adjusting for the chance grouping of elements and it is defined as:

$$ARI = \frac{\text{Rand - Expected (Rand)}}{1 - \text{Expected(Rand)}}, \tag{1.5}$$

where *Expected (Rand)* is the mean of *Rand* given the hypothesis that the two clustering exercises are independent (Qannari, Courcoux, and Faye, 2014). The ARI should be interpreted as follows: ARI >= 0.90 excellent recovery; 0.80 =< ARI < 0.90 good recovery; 0.65 =< ARI < 0.80 moderate recovery; ARI < 0.65 poor recovery (Tellaroli et al., 2018). More details about the ARI can be found in Hubert and Arabie (1985).

1.3.2 Functional Data Analysis

Functional Data Analysis (FDA) is defined as:

> **Definition** (*Functional Data Analysis*).
> Functional Data Analysis extends classical multivariate statistical methods by performing statistical analysis with data represented by curves (i.e., functions) varying over a continuum (Jacques and Preda, 2014a; Jacques and Preda, 2014b).

A functional datum is a continuous function *x(t)* of a variable observed over some interval such as time (Ramsay, 1982). Thus, functional data is represented by a set of curves observed through time instead of discrete data points. These curves belong to a theoretic infinite-dimensional space (Ferraty and Vieu, 2006). However, it difficult to model such data since the observations are supposed to exist in an infinite-dimensional space but in reality one only has curves observed from a finite-dimensional space. Moreover, probabilistic model-based methods do not directly help with clustering due to the lack of a definition for the probability density of a functional random variable (Delaigle and Hall, 2010). For this reason, distances cannot be computed in this context hence distance-based methods are not applicable. However, if curves are represented in a finite-dimensional space then clustering algorithms for finite-dimensional data *can* be utilized. Consequently, reconstruction of data from discrete observations into their functional form is often the first step in FDA. (Jacques and Preda, 2014a).

The concept of FDA was born with Ramsay (1982) where the author acknowledged the work (only available in French) of Cailliez and Pagès (1976) and Dauxois and Pousse (1976) as inspiration for transforming traditional data analysis into the "language of functional analysis." Additionally, a functional datum is described as each observation where subjects are paired with variables and each pair has a data point recorded in time per a data collection experiment.

FDA is further described as the field that "deals with the analysis and theory of data that are in the form of functions, images, and shapes, or more general objects" (Wang, Chiou, and Müller, 2016, p. 2). According to Ramsay and Silverman (2005), the goals of FDA are to showcase data in manners that add value to analysis and underline its attributes, as well as study latent patterns and variation.

For instance, one or more measures can be observed for patients during a finite period. Once converted into their functional form, each of these curves are considered as a single observation which can be summarized as an average curve over a period of time. Therefore, the variation between curves can be measured for comparison (Ramsay, 1982), with methods such as model-based clustering. To recapitulate, one of the main advantages of functional data is that one individual is considered as a curve (or set of curves) and not a vector of points as in multivariate statistics. Thus, functional data analysis allows the dependency between discrete data points in time to be kept.

An example of functional data comes from the CanadianWeather dataset which can be found in the `FDA` package in `R` and contains daily temperature and precipitation data from 35 different

locations in Canada, averaged over 1960 to 1994. Figure 1.4 represents the *mean* temperatures of four Canadian weather stations (Prince Rupert, Montreal, Edmonton, and Resolute) plotted as smoothed curves that we assume generated them.

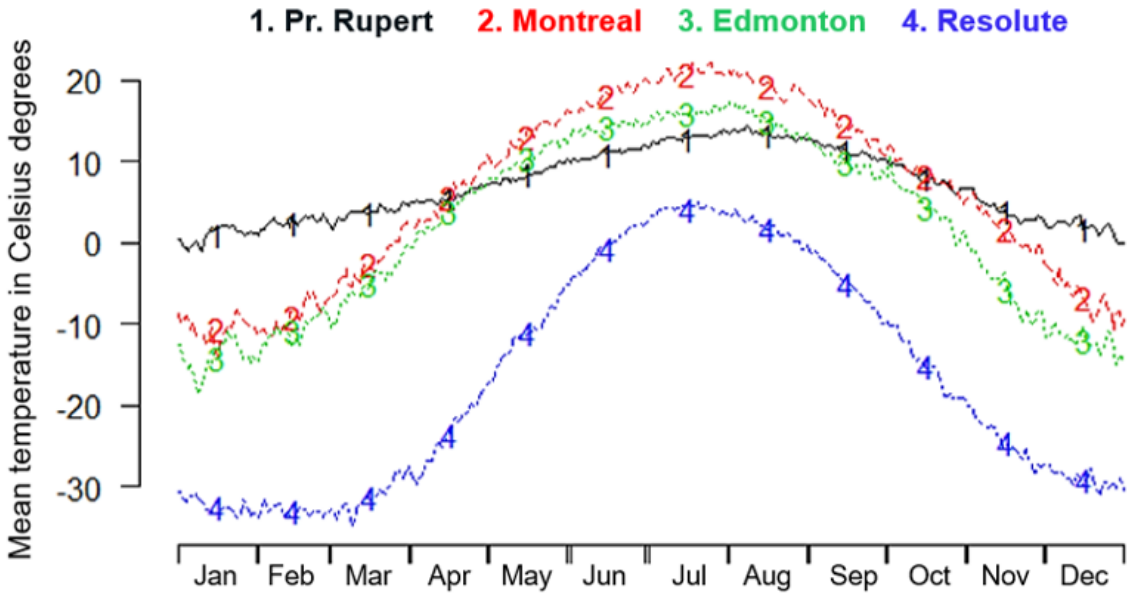


FIGURE 1.4: Mean temperatures at four Canadian weather stations.

Functional data are increasingly prevalent. Instances of functional data can be produced by optical tracking equipment, electrical measurements, astronomy observations, signal processing, and weather data. As mentioned above, these data are of infinite dimension which is why analysis is more arduous and demanding than multivariate or high dimensional data. In the multivariate context, functional data "refers to a set of several functions or times series describing the same individual" (Bouveyron et al., 2019, p. 358).

On the one hand, classical cluster analysis could yield useful for processing multivariate functional data, but the functional nature of the data makes the task difficult as the data live in an infinite-dimensional space. Furthermore, the main disadvantage of the classical approach is that the underlying continuity of the data is ignored, which is essential to accurately represent functional data as we do not want to assume nothing happened between sampled points (Ramsay, 1982). On the other hand, another classical response is to utilize a family of functions to represent the data but this approach is highly inflexible because of the dependency on a limited number of parameters as many datasets are too complex to be fully represented with parametric models. Therefore, the use of piecewise polynomials or *splines* has been a major development as they can use *breaks* to allow when curves need to have sharp peaks or valleys or an abrupt change of level. Breaks in splines make it possible to have more flexibility in representing the data (Hastie, Tibshirani, and Friedman, 2009).

Clustering methods help understand system behavior but come with disadvantages when handling functional data as most clustering methods for functional data apply multivariate techniques to the discretized curves or use distance-based algorithms (Bouveyron et al., 2019). Additional challenges of working with functional data are estimating functional data from noisy or categorical samples, regularization and smoothness, and measures of variation and confidence in estimates (Yao, Müller, and Wang, 2005).

With the ever-increasing computing and storage capacity, many fields have begun to collect and store massive amounts of data in the form of *time series* such as credit card records, stock prices, temperature, and biological measurements. Technically, functional data can be seen as multivariate time series. The main difference between functional data analysis (FDA) and time series analysis (TSA) is in the representation of data points. For FDA, each observation is represented by a curve, whereas for TSA the atomic observation is an individual time point. Therefore, TSA's goal is to analyze temporal dependence between data points and predict *new* time points. Instead, FDA attempts to find common patterns between the curves (i.e., clustering) or response variables in regression models (Ramsay and Silverman, 2005).

Model-Based Clustering for Multivariate Functional Data

There is copious research that present clustering for *univariate* functional data, such as James and Sugar (2003), Bouveyron and Jacques (2011), Jacques and Preda (2013), and Bouveyron, Côme, and Jacques (2015). However, we require an MBCFD algorithm which can handle *multivariate* functional data. A large part of the earlier work on multivariate functional data analysis focused on a variation of k-means, such is the work of Singhal and Seborg (2005) where they modify k-means to cluster multivariate time-series data using similarity factors; Ieva et al. (2013) use multivariate functional k-means with different distance options; Tokushige, Yadohisa, and Inada (2007) extend existing crisp and fuzzy k-means clustering algorithms to the multivariate functional data case; Zambom, Collazos, and Dias (2019) explore hypothesis testing k-means by clustering curves with different degrees of smoothing. Moreover, functional principal component k-means was developed by Yamamoto (2012) with the purpose of seeking "the subspace that is maximally informative about the clustering structure in the data" (p. 1).

However, the latest developments in FDA have presented efficient model-based clustering methods for functional data. Model-based clustering is defined as:

> **Definition** (*Model-based Clustering*).
> "Model-based clustering is a principled approach to cluster analysis, based on a probability model and using standard methods of statistical inference. The probability model on which it is based is a finite mixture of multivariate distributions" (Bouveyron et al., 2019, p. 15).

Model-based clustering involves the following elements: (1) model: based on a finite mixture probability model, (2) estimation method: systematic statistical method for model parameters estimation, and (3) method for classification: systematic method for classifying the observations conditionally on the model (Bouveyron et al., 2019). Mixture modeling assumes that the data is sampled from a population described by a probability density function which is "characterized by a parameterized model taken to be a mixture of component density functions; each component density describes one of the clusters" (Hastie, Tibshirani, and Friedman, 2009).

A model-based approach utilizes particular models for clustering and attempts to optimize the fit between the data and the model by maximum likelihood or Bayesian approaches. Recent breakthroughs in functional data analysis have facilitated the development of model-based techniques for clustering functional data. We adopted the classification of functional data clustering methods and their respective definitions (see Figure 1.5) from Jacques and Preda (2014a):

- Raw data methods: these methods consist of clustering directly the curves on the basis of their evaluation points.

- Distance-based methods: these methods use clustering algorithms based on specific distances for functional data. Notice that, depending on the way these distances are computed, these methods can be related to either raw data or filtering methods.
- Filtering methods: these methods first approximate the curves into some basis of functions and second perform clustering using the basis expansion coefficients.
- Adaptive methods: these methods consider that the functional representation of data is depending on clusters, and perform simultaneously dimensionality reduction and clustering. Thus, depending on its cluster membership, an observation (a curve) could have different representations.

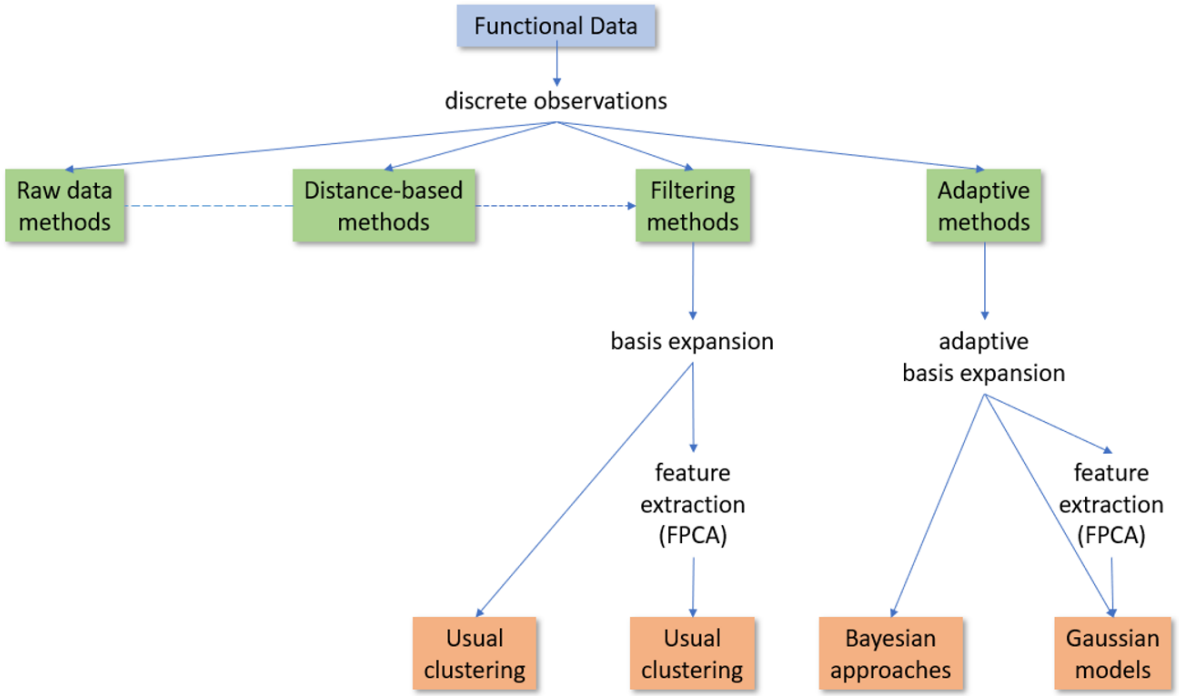


FIGURE 1.5: Segmentation of different clustering methods for functional data adapted from Jacques and Preda (2014a), p. 238.

1.3.3 SQ1.2: Which are MBCFD algorithms that can be used for deriving knee OA phenotypes?

To answer our research question, we focused on the adaptive methods category. Adaptive methods are model-based clustering methods for functional data. In this category, the basis expansion coefficients are treated as random variables with cluster-specific probability distribution instead of parameters as in filtering methods. Put differently, the probability of cluster membership can be estimated for each observation for all clusters. Most adaptive methods use probabilistic clustering of basis expansion coefficients or functional principal component analysis scores.

We focused our options on the two most recent adaptive model-based clustering methods: funFEM and funHDDC. Functional PCA and latent mixture models are used as well as model estimation by Expectation-Maximization (EM) in both methods. We evaluated funFEM and funHDDC based on flexibility, interpretability and ability to handle multivariate data. We define

these criteria as: (i) *flexibility*: ability of the method to adjust its parameters, (ii) *interpretability*: access to model results and possibility to easily visualize resulting clusters, (iii) *ability to handle multivariate data*: option to use univariate and multivariate data.

funFEM. This model was developed by Bouveyron, Côme, and Jacques (2015) and is based on a discriminative functional mixture (DFM) model. funFEM models variables within each discriminative functional subspace separately. The goal is to produce a straightforward visualization of the clusters to be able to compare the discovered patterns. funFEM can be considered an extension of discriminative latent mixture (DLM) models into the functional case. In DLM models, the latent subspace is the one that maximizes the separation between clusters hence common to all groups. In funFEM, the data is represented via basis functions, then the most discriminative space is found and the maximum likelihood is obtained with an EM algorithm.

funFEM is well-equipped in terms of flexibility and interpretability, however, it is not yet designed to handle multivariate functional objects.

funHDDC. Clustering in high-dimensional spaces is complex because high-dimensional data typically live in different low-dimensional subspaces hidden in the original space. Bouveyron, Girard, and Schmid (2007) developed a method based on the Expectation-Maximization algorithm called High-Dimensional Data Clustering (HDDC) which estimates the *specific* subspace and the intrinsic dimension of each group. Bouveyron and Jacques (2011) expanded HDDC to the functional case with funHDDC which assumes a parsimonious cluster-specific Gaussian distribution for basis expansion coefficients. Then, Jacques and Preda (2014b) proposed Funclust, the first model-based clustering algorithm for *multivariate* functional data. Funclust is a Gaussian mixture model based on multivariate functional principal component analysis (MFPCA), defined and estimated by a variation of the EM algorithm. The main advantage of this method is that the dependency between the curves is included in the analysis through MFPCA. Funclust is quite flexible given its probabilistic modeling but it has the limitation of only modeling a proportion of the variance and thus part of the information is missed. Furthermore, Schmutz et al. (2020) adapted funHDDC to the multivariate case by using a "functional latent mixture model which fits the data into group-specific functional subspaces through a multivariate functional principal component analysis" (Schmutz et al., 2020, p. 1). funHDDC overcomes the limitation of Funclust by modeling all non-null variance principal components thus all information is considered. Moreover, the funHDDC method assumes that the scores of the functional principal components have Gaussian-distributed and cluster-specific parameters and uses Expectation-Maximization to infer/select the best model. The choice of hyper-parameters is achieved through embedded model selection.

Consequently, we selected **funHDDC** as our MBCFD method to derive knee OA phenotypes due to its flexibility, interpretability, and ability to handle multivariate data. Up to now, only Funclust (Jacques and Preda, 2014b) and funHDDC are able to handle multivariate data (Bouveyron et al., 2019). With funHDDC being built upon Funclust, it would be redundant to consider both methods.

funHDDC is flexible since it allows us to select between six levels of model parsimony, adjust the number of EM iterations, test several numbers of clusters at the same time, choose from three criteria for model selection (BIC, ICL and slope heuristic), among other parameters. The interpretability of the results when plotted as line charts allow us to see three dimensions, namely time, trajectory (i.e., direction or slope) and measurement. In addition, the results of funHDDC show the number of dimensions for each cluster, the mean of each cluster in the original space, the proportion of individuals in each cluster, the maximum log-likelihood, the log-likelihood at each iteration, the posterior probability for each individual to belong to each cluster, the clustering partitions, BIC score, ICL score, and number of parameters estimated.

funHDDC uses *basis functions*, which are like functional building blocks, to model the variables through time by reconstructing them into curves. One of the advantages of using a basis function system is that we do not take noise into account. The linear combination of the *basis functions* is called a *basis function expansion*. In the same way, a *basis system* is a linear combination of the monomial *basis functions* with their respective coefficients. Since polynomials are limited in their flexibility in modeling complex functional shapes, Fourier series and splines are widely used to overcome this limitation. Thus, a *basis system* can be modeled via Fourier series for periodic data and b-splines for non-periodic data. These two systems are supplemented by the monomial and constant *basis systems*. Fourier series and b-splines can usually handle the vast amount of analysis problems, though there are other systems available such as the exponential basis for exponential functions, polygonal basis for straight-line segments, and the power basis for a sequence of noninteger powers of an argument t (Ramsay, Hooker, and Graves, 2009).

Multivariate functional principal component analysis (MFPCA) was first proposed in Ramsay and Silverman (2005) as an extension of PCA for functional data to the multivariate case by concatenating the coefficients in a basis expansion into a vector and then performing regular PCA on the concatenated vectors. However, Schmutz et al. (2020) sustain that this approach has the limitation of modeling only a proportion of principal components, therefore, missing a considerable portion of the information. For this reason, the funHDDC algorithm by Schmutz et al. (2020) overcomes this obstacle by modeling all principal components with non-null variance resulting in an improved clustering exercise based on a Gaussian model-based clustering method which manages the dependency between functional variables.

Chapter 2

Problem Statement

In data mining problems, a patient can be represented by their Electronic Health Records (EHR). In order to find knee OA phenotypes, we can use EHR data to create groups that are similar within and different between each other. EHR data can be categorized as functional data as it is evidenced in the form of quantitative measurements evolving through time. In the univariate case, functional data X is represented by a single curve where t represents the time:

$$X(t)\,\varepsilon\,\mathbb{R}\,\forall t\,\varepsilon\,[0,T] \tag{2.1}$$

A patient is typically represented by more than one characteristic and data on these characteristics is collected in a clinical, trial or laboratory setting. Therefore, the corresponding multivariate functional data, or set of n p-variate curves, can be written as:

$$\mathbf{X} = \mathbf{X}(t)_{t\varepsilon[0,T]}\ with\ \mathbf{X}(t) = (X^1(t), ..., X^p(t))'\,\varepsilon\,\mathbb{R}^p,\ p \geq 2 \tag{2.2}$$

In reality, the functional expressions of the curves are not known and it is only feasible to have discrete observations at a finite vector of times available. Consequently, it would be required to convert these discretely observed values into a function $X_j^i(t)$ computable for the desired time argument $t\,\varepsilon\,[0,T]$.

Table 2.1 shows the terminology used, i.e., the variables and their meaning.

In other words, functional data can be seen as multivariate data with order (a continuum) in its dimensions. Therefore, patients' observations through time could be considered as functions because each curve is the observation of a measure/variable over a finite period of time. After all, time is a crucial dimension to consider because it could allow us to identify more complex disease patterns and often goes ignored in similar analyses. Functional data can keep the time dependency between data points by converting vectors of discrete data points into curves thus allowing information between sampling points to remain in the analysis.

Consequently, we aim to determine whether an MBCFD method improves the discovery of phenotypes versus a non-functional, more traditional clustering approach, namely HCA. The technical problem can be stated as:

> **Problem Statement.**
>
> Given a set of objects $\mathbf{X}$ in an p-dimensional space, we want to identify K clusters with a model-based clustering of functional data algorithm that outperforms hierarchical cluster analysis in terms of clinical relevance and statistical significance.

Our research clusters structured data features with the expected output in the form of statistically-significant and clinically-relevant subgroups. In order to derive phenotypes from structured data, we specifically investigate using an MBCFD method because of its flexibility in interpretation, its ability to maintain the continuity information between multivariate sampling points, and to assess whether a functional data clustering method performs better at identifying phenotypes than hierarchical cluster analysis, a very commonly used algorithm for detecting phenotypes. However, the only way to compare unsupervised clustering algorithms is to apply them

TABLE 2.1: Terminology used to describe univariate/multivariate functional data.

Variable	**Meaning**
X	Data represented by curves
t	Time
T	Endpoint of time interval
p	Number of dimensions
n	Number of curves

on data with ground truth labels so that we are able to estimate the percentage of error for each algorithm. Given that we do not have access to data with ground truth labels, we can focus on statistical significance and clinical relevance between the groups of both exercises.

In summary, we expect to find statistically-significant, clinically-relevant subgroups in the data that differ in disease course (10-year trajectories) and etiological mechanisms (features) through the use of an MBCFD algorithm that outperforms a traditional clustering approach.

Chapter 3

Solution

This experiment was conducted at the Rheumatology and Clinical Immunology Department of the University Medical Center Utrecht (UMCU) in The Netherlands. As a university medical center, the UMCU is tasked with performing research, which makes it appropriate for conducting this experiment. The dataset used for analysis is from the Cohort Hip and Cohort Knee (CHECK) study. Medical experts and researchers familiar with CHECK, OA etiology and statistical analysis were consulted throughout the process. The method used for the experiment is CRISP-IDM which is described in a subsequent section and in detail throughout this chapter.

3.1 Data: Cohort Hip and Cohort Knee (CHECK)

CHECK, an initiative of the Dutch Arthritis Foundation, is a multi-center 10-year prospective cohort study of 1002 individuals with signs of early symptomatic OA of hip or knee, aged 45-65 years, and without a previous consultation for these complaints or with a first consultation no longer than six months ago to their primary care physician. Participants who potentially fulfilled the inclusion criteria were asked to join the study when visiting their physicians. Moreover, participants were recruited through newspapers, ads, and through the Dutch Arthritis Foundation[1] web site. The participating centers' medical ethics committees approved the study and all participants signed informed consent forms. Patients with other conditions that could explain their symptoms were excluded. Other exclusion criteria were comorbidities that impeded physical evaluation or 10-year follow up, presence of malignancy in the last five years and Dutch language proficiency (Wesseling et al., 2014). Participants were evaluated clinically through regular examinations and questionnaires, radiographically via knee and hip radiographs, and biochemically with collection of plasma, serum, and urine samples (Van Spil et al., 2012). Participants were divided into two groups: annual and variable, depending on the severity of their symptoms with patient with more serious symptoms visiting the centers each year and patients with milder symptoms visiting at years 0, 2, 5, 8 and 10. The aim of the study was to help improve knowledge regarding early OA.

CHECK data were collected at ten general and university hospitals in The Netherlands, located in semi-urbanized regions. The participating centers were Erasmus Medical Center Rotterdam, Kennemer Gasthuis Haarlem, Leiden University Medical Center, Maastricht University Medical Center, Martini Hospital Groningen/Allied Health Care Center for Rheumatology and Rehabilitation Groningen, Medical Spectrum Twente Enschede/Ziekenhuisgroep Twente, Reade, Center for Rehabilitation and Rheumatology, St Maartenskliniek Nijmegen, University Medical Center Utrecht, and Wilhelmina Hospital Assen. The participating centers' medical ethics committees approved the study and all participants signed informed consent forms (Wesseling et al., 2014).

[1] https://reumanederland.nl/

3.2 CRISP-IDM Method

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) method was conceived in 1996 by DaimlerChrysler, SPSS, and NCR (Chapman et al., 2000). CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. CRISP-DM is broadly considered to be one of the best knowledge discovery methods in data mining, in part because it seamlessly includes organizational aspects of data mining. The method chosen to perform the OA phenotype research, CRISP-IDM (Menger et al., 2016), is a specification of the CRISP-DM method applicable to exploratory and interactive data analysis in healthcare and potentially other fields. Menger et al. (2016) adapted the CRISP-DM method with the goal in mind to do exploratory data analysis that integrates domain experts (particularly in the healthcare industry), and created the CRoss-Industry Standard Process for *Interactive* Data Mining (CRISP-IDM) method (see Figure 3.1) by introducing three modifications: 1) aggregating the modeling and evaluation phases into one iterative phase, which requires the involvement of domain experts; 2) separating the data preparation phase into general and specific. The general data preparation will be part of the modeling phase and the specific data preparation will be part of the evaluation phase, and 3) adding an optional inferential analysis step necessary to bring exploratory analysis results and/or generated hypotheses into practice with sufficient statistical confidence.

A brief overview of the phases is as follows, more in-depth information is provided in the subsequent sections.

1. Domain understanding: the initial phase consists of understanding project objectives and requirements through the organization of topics, themes and their priorities by interviewing domain experts in context.

2. Data understanding: starts with initial data collection and follows with activities that enable familiarity with the data, identifying data quality issues, and noticing some insights. This phase requires the selection of relevant internal and external data sources such as Electronic Health Records (EHR), imaging data, laboratory results, census data, etc.

3. General data preparation: relates to the preprocessing of the data to convert it into the appropriate format for exploratory analysis. The most relevant tasks in preparing the data are transforming, cleaning, integrating, reducing, and discretizing the data (Zhang, Zhang, and Yang, 2003).

4. Modeling and evaluation: an interactive data visualization tool is used to involve domain experts in modeling the data, which enables immediate feedback. This phase consists of five iterative activities (see Figure 3.1). The double-sided arrow between the specific data preparation and set up visualization activities denote the interactive back-and-forth that takes place inside this phase.

5. Inferential analysis: this phase is optional (depends on the project's analysis needs and data available) and aids in implementing exploratory results or generated hypotheses with sufficient statistical confidence.

6. Deployment: focuses on the implementation of results transforming them into daily practice.

The following sections provide an in-depth description of the execution of the experiment per each of CRISP-IDM's phases.

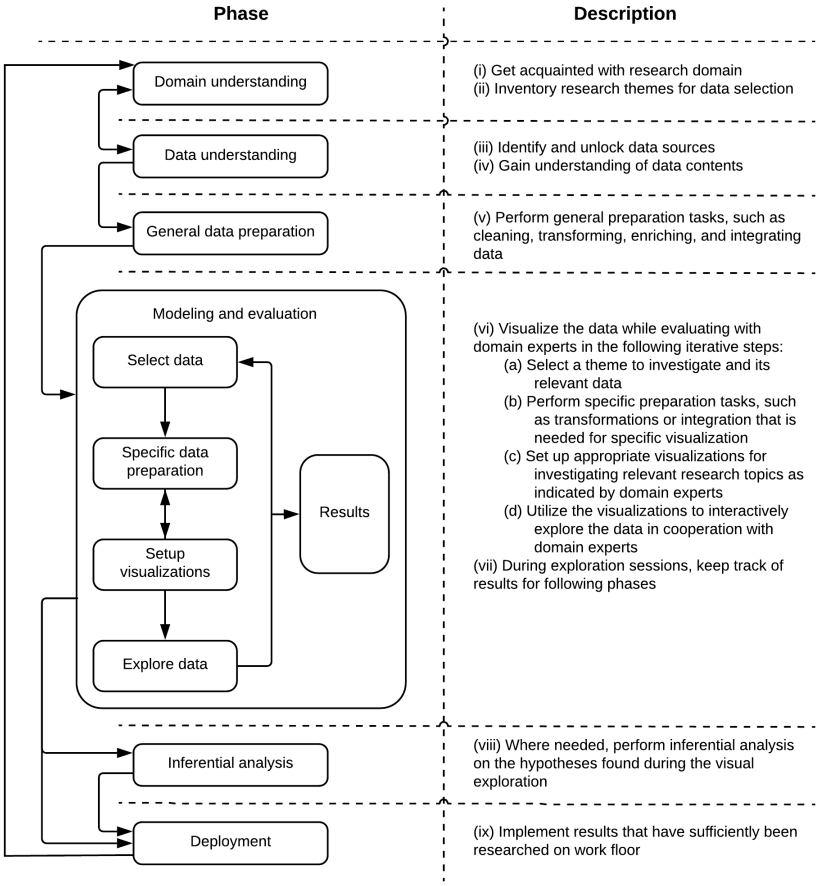


FIGURE 3.1: Overview of the CRISP-IDM method, adapted from Menger et al. (2016, p. 3).

3.3 Domain Understanding

In order to become acquainted with the objectives and requirements of the project, two UMCU medical experts were regularly consulted, as well as the main supervisor of this work from Utrecht University. Additionally, two statistical analysis experts were consulted at the beginning and at the end of the project. Meetings with medical experts allowed the mapping of themes and topics. Themes are needed for selecting relevant data sources and general data preparation. A total of seven topics were identified after the meetings with experts, which can be seen in Table 3.1. These seven topics were grouped into three themes, namely 'state of the art', 'method performance', and 'relevant phenotypes'. The topics are also the basis for defining our research questions which can be seen in the Introduction section of this thesis. The prioritization was determined by the contribution of each topic to the MRQ. The 'state of the art' theme relates to the current research being performed to discover OA phenotypes and requires data from the literature review. The 'method performance' theme relates to how well the selected MBCFD method performs at identifying phenotypes from the CHECK data and in comparison with a widely-used method in the scientific literature. Lastly, the 'relevant phenotypes' theme pertains to detecting statistically-significant and clinically-relevant phenotypes. The second and third themes require data from CHECK which is further explained in the Data Understanding phase.

TABLE 3.1: The seven topics identified during the domain understanding phase along with their corresponding theme and priority.

Topic	**Theme**	**Priority**
Current and most common methods for discovering knee OA phenotypes	State of the art	2
Which characteristics (groups of features) are used for deriving knee OA phenotypes?	State of the art	2
Which Model-based Clustering of Functional Data (MBCFD) algorithms are suitable for the analysis?	State of the art	1
Performance of selected MBCFD algorithm	Method performance	1
Comparison of MBCFD algorithm with method widely used in the literature	Method performance	2
Detect statistically-significant phenotypes	Relevant phenotypes	1
Detect clinically-relevant phenotypes	Relevant phenotypes	1

Clinically-Relevant Phenotypes. We define clinically-relevant phenotypes as those that are represented by different trajectories between features and/or present upward or downward trajectories. In addition, it is also considered interesting when two or more phenotypes show progression in one feature but have different progression in another feature. We define *progression* as trajectories with either upward or downward trends. According to domain expertise, we are most interested in finding increasing or decreasing progression over time and the synergistic effect when combining features, e.g., when a lateral feature increases over time but its medial counterpart decreases or remains the same. Figure 3.2 shows examples of what could be deemed as clinically-relevant phenotypes by domain experts for multivariate scenarios. Moreover, it is essential to maintain the number of clusters low so phenotypes can be described and used in further research. Lastly, it would not practical nor relevant to have a large number of clusters, however, we need enough clusters to represent progression. The least relevant phenotypes are ones where we see either high, moderate, or low constant levels (i.e., straight trajectories with no upward or downward trends). Additionally, since we are interested in finding inter-feature differences, the same behavior (i.e., trend) for all features is deemed as not interesting.

The clinical relevance criteria can be summarized as:

(i) Different inter-cluster feature trajectories

(ii) Upward/downward trajectories

(iii) Multi-feature different behavior

(iv) Balanced number of clusters

3.4 Data Understanding

Each of the previously identified themes requires selecting and accessing relevant data sources. Relevant English-language articles will be used as the data source for the 'state of the art' theme and presented in chapter 1 of this thesis. However, the literature review is not considered input for the technical analysis, thus it is not be reflected in Table 3.2. In the case of the

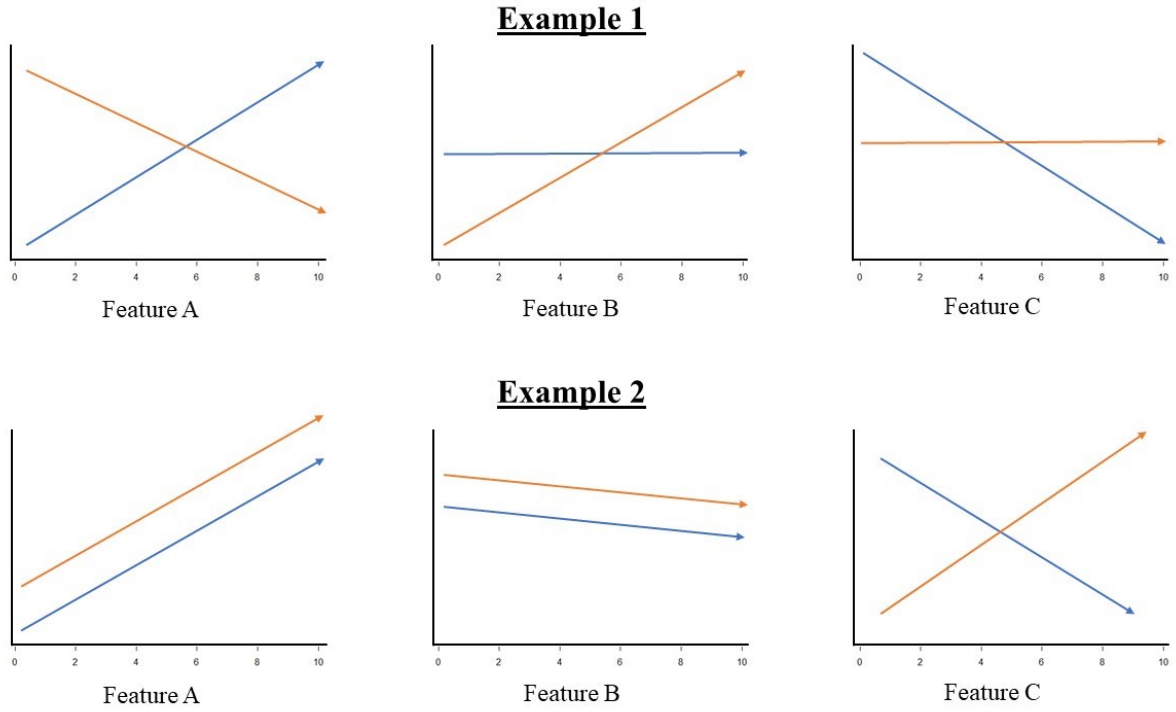


FIGURE 3.2: Examples of trajectories of clinically-relevant phenotypes.

'method performance' and 'relevant phenotypes' themes, we are using the CHECK dataset which is described on the CHECK Research website and access can be requested through the online archiving system EASY of Data Archiving and Networked Services (DANS). The data is provided in comma-separated values (CSV) format. Even though a subset of CHECK data was collected at the UMCU, other medical centers participated in the study, therefore CHECK data is considered an *external* source. All acquired data entities and their type, structure, and number of records are listed in Table 3.2 and described below. Baseline characteristics for CHECK participants are shown in Table 3.3. Additionally, Tables 3.4, 3.5, 3.6, and 3.7 present a summary of collected data during 10 years in all participants (A) and in the subgroup (S) of the annual visiting group of CHECK for the four different categories: questionnaires, clinical assessment, radiographic assessment and biochemical markers. Subjects with mild symptoms visited the medical centers at years 0, 2, 5, 8 and 10 and subjects with significant symptoms visited the medical centers annually. For an in-depth description of the CHECK study, see Wesseling et al. (2014).

TABLE 3.2: Acquired data entities with type, structuredness, and number of records.

Data entity	**Type**	**Structured/ unstructured**	**Number of records**
Questionnaires	Continuous, discrete	Structured	11,022
Clinical assessment	Continuous, discrete	Structured	11,022
Radiographic assessment	Continuous, discrete	Structured	5,010
Biochemical markers	Continuous, discrete	Structured	1,002

TABLE 3.3: Baseline characteristics of CHECK participants.

Characteristic	**N(%)**
Age in years, mean (sd)	56 (5)
Females, n (%)	792 (79)
Post menopausal, n (%)	475 (77)
Caucasian, n (%)	976 (98)
BMI in kg/m2, mean (sd)	26 (4)
Academic or higher vocational education, n (%)	267 (28)
Physical activity (more than 30 minutes) for three times a week or more, n (%)	524 (54)
Smoking every day, n (%)	90 (9)

N: number of subjects; sd: standard deviation

3.4.1 Questionnaires and Clinical assessment

Clinical variables in the CHECK study were collected via self-reported questionnaires, medical history questions and physical examination by a health practitioner. Self-reported questionnaires evaluated hip and knee symptoms, hand symptoms, pain severity, coping, health-related quality of life, leisure activities and employment, economic consequences, social support and comorbidities (Wesseling et al., 2014). Summaries for the questionnaires and clinical assessment data can be seen in Tables 3.4 and 3.5.

TABLE 3.4: Summary of collected questionnaires data from CHECK, adapted from Wesseling et al. (2014).

Questionnaires data/year	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Demographics	A	S	A	S	S	A	S	S	A	S	A
SF-36: Short-Form 36-item health status survey	A	S	A	S	S	A	S	S	A	S	A
EQ5D: EuroQol (Quality of Life)	A	S	A	S	S	A	S	S	A	S	A
WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index	A	S	A	S	S	A	S	S	A	S	A
NRS for pain intensity (numerical rating scale)	A	S	A	S	S	A	S	S	A	S	A
Comorbidity list	A	S	A	S	S	A	S	S	A	S	A
Health care use	A	S	A	S	S	A	S	S	A	S	A
Pain Coping Inventory list	A		A			A			A		A
Social Support scale	A		A			A			A		A
Lifestyle: tobacco and alcohol use	A		A			A			A		A
AUSCAN: Australian Canadian Osteoarthritis Hand Index									A		
ICOAP: Measure of Intermittent and Constant Osteoarthritis Pain									A	S	A

A: all participants; S: subgroup or annual visiting group

TABLE 3.5: Summary of collected clinical assessment data from CHECK, adapted from Wesseling et al. (2014).

Clinical assessment data/year	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Knee examination											
Palpable warmth	A	S	A	S	S	A	S	S	A	S	A
Refill test	A	S	A	S	S	A	S	S	A	S	A
Bony tenderness	A	S	A	S	S	A	S	S	A	S	A
Patella grinding test	A	S	A	S	S	A	S	S	A	S	A
Range of motion: flexion/extension	A	S	A	S	S	A	S	S	A	S	A
Crepitus	A	S	A	S	S	A	S	S	A	S	A
Hip examination											
Range of motion: flexion/internal/ external rotation/adduction/abduction	A	S	A	S	S	A	S	S	A	S	A
Hand examination											
DIP/PIP bony enlargements	A	S	A	S	S	A	S	S	A	S	A
CMC I bony enlargements							S	S	A	S	A
Soft tissue swelling MCP I-V							S	S	A	S	A
Deformity CMC I, DIP, PIP							S	S	A	S	A

A: all participants; S: subgroup or annual visiting group

3.4.2 Radiographic Data

The severity of knee and hip osteoarthritis is scored according to the Kellgren and Lawrence (K&L) scale (0-4) with grade 0 signifying no presence of OA and grade 4 signifying severe OA (Kellgren and Lawrence, 1957). Separate features of the knee and hip were scored on other radiographs according to Altman and Gold (2007) and Burnett et al. (1994), both on a 0–3 scale. The radiographs were independently scored by five experienced observers. Readers scored all consecutive radiographs simultaneously with a known sequence, but blinded to the clinical status. Interobserver variability was tested resulting in moderate to substantial interobserver agreement (Wesseling et al., 2014). Lastly, Knee Images Digital Analysis (KIDA) assessed more comprehensive quantitative parameters on radiographs (Marijnissen et al., 2008) and were measured without knowing the sequence of the radiograph. A summary of collected radiographic assessment data can be seen in Table 3.6.

TABLE 3.6: Summary of collected radiographic assessment data from CHECK, adapted from Wesseling et al. (2014).

Radiographic assessment data/year	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Knee: unilateral posterior-anterior fixed exion view (both knees)	A		A			A			A		A
Knee: unilateral lateral view (both knees)	A		A			A			A		A
Knee: bilateral skyline view (supine)	A		A			A			A		
Hip: anterior-posterior pelvis view	A		A			A			A		A
Hip: unilateral faux profile view (both hips)	A		A			A					
Hand: bilateral posterior-anterior view									A		
Lumbar spine: lateral view (supine)									A		

Table 3.6 continued from previous page

Radiographic assessment data/year	0	1	2	3	4	5	6	7	8	9	10

A: all participants; S: subgroup or annual visiting group

3.4.3 Biochemical Markers Data

Markers of cartilage, bone and synovial metabolism were collected to enhance understanding of mechanisms of progression. Blood and urine samples were acquired from each subject adhering to a standardized protocol at all participating centers. A systematic review of the biochemical markers in knee and hip OA was the foundation for formulating the set of biochemical markers to be garnered at baseline (Van Spil et al., 2010). A summary of collected biochemical assessment data can be seen in Table 3.7.

TABLE 3.7: Summary of collected biochemical assessment data from CHECK, adapted from Wesseling et al. (2014).

Biochemical assessment data/year	0	1	2	3	4	5	6	7	8	9	10
DNA	A										
Plasma	A		A			A					
Serum	A		A			A					
Urine	A		A			A					
Biochemical markers	A										

A: all participants

3.5 General Data Preparation

The data was provided in CSV format as a long dataset with 573 columns (variables) and 11,022 rows. Each subject has been anonymized by means of a subject identification number (NSIN). All data processing was performed in the Rstudio Integrated Development Environment (IDE) with the R language (version 3.6.2) (R Core Team, 2013). RStudio includes a console, syntax-highlighting editor, diverse support tools and workspace management capacity. `R` is an open source language widely used for statistical analysis.

In order to convert all variables to the appropriate format for exploratory data analysis, the following actions were taken:

1. Checked the minimum, maximum, range, mean/median, standard deviation, and number of missing values for all variables to identify inconsistencies.

2. Converted variables to appropriate data types with functions such as as.factor(), as.integer(), and as.numeric() from the `base` package.

3. Replaced characters such as commas with dots (e.g., "5.5" instead of "5,5") with the pattern matching and replacement function gsub() from the `base` package.

4. Removed negative and out-of-scale values where appropriate (e.g., when measuring joint space narrowing distances, negative values are considered errors, thus removed from the dataset). These outliers were most likely generated due to human error.

5. Converted binary variables from 1/2 to 0/1

Testing Normality. During the fifth phase of CRISP-IDM, Inferential Statistics, we performed hypothesis testing to understand whether the derived clusters have a statistically-significant difference in inter-cluster means. For this purpose, we visually inspected the data by creating histograms of variables per year and applied the Shapiro-Wilk test to the data to learn if it is normally-distributed. We used the `shapiro.test` function from the `stats` package (R Core Team, 2019) in `R`. As we can see from the results in Table 3.8, the data is not normally distributed as the p-values < 0.05, hence, indicating the data considerably deviate from a normal distribution. Thus, given the data is not normally distributed, it is recommended to use the Kruskal-Wallis test to ascertain statistical significance.

TABLE 3.8: Shapiro-Wilk test results for normality.

Data/p-values	WOMAC Pain	WOMAC Function	WOMAC Stiffness	KIDA	OA Scoring
Baseline	6.54E-15	1.86E-16	2.79E-15	3.99E-08	1.67E-68
Year 1	3.84E-16	4.28E-18	6.96E-16	2.77E-11	4.66E-68
Year 2	6.49E-17	1.16E-18	1.11E-16	6.95E-11	5.63E-68
Year 3	6.52E-19	8.32E-21	2.39E-17	3.29E-08	4.33E-67
Year 4	4.36E-18	1.14E-19	2.97E-17	4.69E-07	4.33E-67
Year 5	3.54E-18	1.08E-17	3.41E-16	4.46E-10	1.00E-59
Year 6	1.98E-17	2.57E-18	3.13E-18	9.31E-14	2.19E-56
Year 7	1.04E-18	3.54E-19	5.83E-18	7.40E-18	2.56E-51
Year 8	8.00E-20	1.20E-19	2.21E-19	1.91E-18	6.03E-51
Year 9	1.61E-20	2.27E-19	2.10E-19	8.90E-20	6.03E-51
Year 10	1.21E-19	2.36E-19	9.50E-19	2.58E-44	4.07E-44

Clustering Tendency. We aimed to learn whether the data might contain meaningful insights. One way to do this is by understanding if the data were generated from a uniform distribution (i.e., meaningful results would not be found). For this purpose, we used the `get_clust_tendency` function from the `factoextra` package (Kassambara and Mundt, 2020) in `R` to compute the H statistic with sample size $n = 100$. As we can see from table 3.9, the results suggest the data does not contain uniformly-distributed data as $H > 0.5$.

In addition, with the `graph` argument from the `get_clust_tendency` function set to `TRUE`, we plotted the dissimilarity matrix based on Euclidean distance and reordered the data points resulting in a ordered dissimilarity matrix. The ordered dissimilarity images can be visualized in Appendix D, Figure D.1. By assessing the plots, we can detect the clustering tendency by counting the number of square-shaped dark blocks along the diagonal (Kassambara and Mundt, 2020). The darker the squares, the more well-separated the clusters (Bezdek and Hathaway, 2002).

TABLE 3.9: Hopkins statistic test results.

Sample data	WOMAC Pain	WOMAC Function	WOMAC Stiffness	KIDA	OA Scoring
n=20	0.69	0.73	0.67	0.82	0.80
n=50	0.73	0.76	0.69	0.79	0.81
n=100	0.72	0.76	0.67	0.81	0.81
n=200	0.72	0.75	0.67	0.80	0.80
Average	*0.71*	*0.75*	*0.68*	*0.80*	*0.81*

The outcome of the General Data Preparation phase was a clean and understandable dataset

ready for analysis. At this stage, only general data preparation requirements were clear to the data scientist, thus *specific* data preparation tasks were executed during the specific data preparation iterations in the modeling and evaluation phase.

3.6 Modeling and Evaluation

The modeling and evaluation phase encompasses the following five interactive activities: select data, specific data preparation, setup visualizations, explore data and results. Since our investigation is explorative in nature, the close collaboration of the data scientist and domain experts is executed throughout the iterative activities of this phase in weekly meetings. The domain experts are needed to guide the analysis with their domain expertise by, for instance, recommending subsets and combinations of features that make clinical sense. Subsequently, the data scientist modifies the `R` scripts or creates new scripts depending on the results of the last activity of this phase. Once the scripts are finalized, the data scientist proceeds to create a document with the visualizations and pertinent information to discuss with the domain experts during the next iteration. We explain in further details activities two through four of this phase in the next sections, as we only have one main data source we do not have to change this selection (first activity). We dedicate separate subsections for the specific data preparation for each algorithm.

3.6.1 Specific Data Preparation

Specific Data Preparation for MBCFD

This sub-activity consists of preparing the functional data objects (i.e., converting discrete observations into curves), specifying the basis system, and running the `funHDDC` algorithm.

Preparing the functional data. The following `R` packages were used: `dplyr` by Wickham et al. (2020), `reshape` by Wickham (2007), `base` included in `R`, `imputeTS` by Moritz and Bartz-Beielstein (2017), `fda` by Ramsay et al. (2018), and `funHDDC` by Schmutz, Jacques, and Bouveyron (2019). Table 3.10 shows the packages, functions and their specific utilization/steps. The `fda` package contains the object class `fd` which is used to represent functional data objects as a finite linear combination of basis functions by creating a `list` of class `fd` which stores the basis functions and individual coefficients for each curve (Happ-Kurz, 2020; Ramsay et al., 2018). Moreover, Listings 3.1 and E.1 respectively show the pseudocode and R code of an example of the data preparation for radiographic variable: left knee sky view patellofemoral osteophyte. Figure 3.3 presents how the WOMAC Pain data is visualized after it has been reconstructed into its functional form. Once the functional data objects have been created, the basis system needs to be specified and the `funHDDC` algorithm needs to be run.

TABLE 3.10: Specific data preparation packages and functions used for the model-based clustering of functional data algorithm.

Package	Function	Utilization/steps
dplyr	select	Select specific variables to create an individual dataframe for each variable which contains three dimensions: the subject identification number (NSIN), the year and the corresponding continuous or discrete value. This step results in a long dataframe with three columns.

Table 3.10 continued from previous page

Package	Function	Utilization
reshape	cast	Cast the long dataframe into the reshaped or aggregated form we desire: a wide dataframe where each column is one year and each row is a patient. Depending on data availability, we may have one, five, and up to eleven years of data.
base	is.na	Subset rows with a maximum of three NAs per row.
imputeTS	na_interpolation	Use linear interpolation to replace remaining missing values per row.
base	intersect	Create an index of the intersection of the prepared dataframes of all variables.
fda	smooth.basis	Construct a functional data object by smoothing data using a roughness penalty. We construct one functional data object per variable. This allows for the combination of a set of coefficients with the specified basis system.
funHDDC	funHDDC	Cluster multivariate functional data into group-specific functional subspaces.

```
-Select year, id, variable from main dataframe
-Save into new functional dataframe for variable
-Cast the functional dataframe into rows for observations and
columns for each year
-Select rows from functional dataframe that have a maximum of
three missing values
-Perform rowwise linear imputation for the functional dataframe
-Save into new functional dataframe for variable
-Create index containing the list of observations
-Intersect index with other variables to create multivariate
index
-Subset the functional dataframe with multivariate index
-Create functional data object with (previously created) basis
system
```

LISTING 3.1: Pseudocode for specific data preparation.

Specifying the Basis System. As stated in chapter 1, there are a few options for basis systems. Recall that a basis system is a linear combination of the monomial basis functions with the coefficients from the functional data. Examples of basis systems are Fourier series, b-splines, exponential, power, and polygonal bases. Figure 3.4 illustrates the *shape* the data would take after *smoothing* them into the chosen basis system. After careful consideration, we decided that exponential (Figure 3.4.D.) and power (Figure 3.4.E.) bases do not apply in our case because we do not have exponential functions nor a sequence of powers as our data. That left us with the choice of Fourier series (Figure 3.4.A.), b-splines (Figure 3.4.B.) or polygonal (Figure

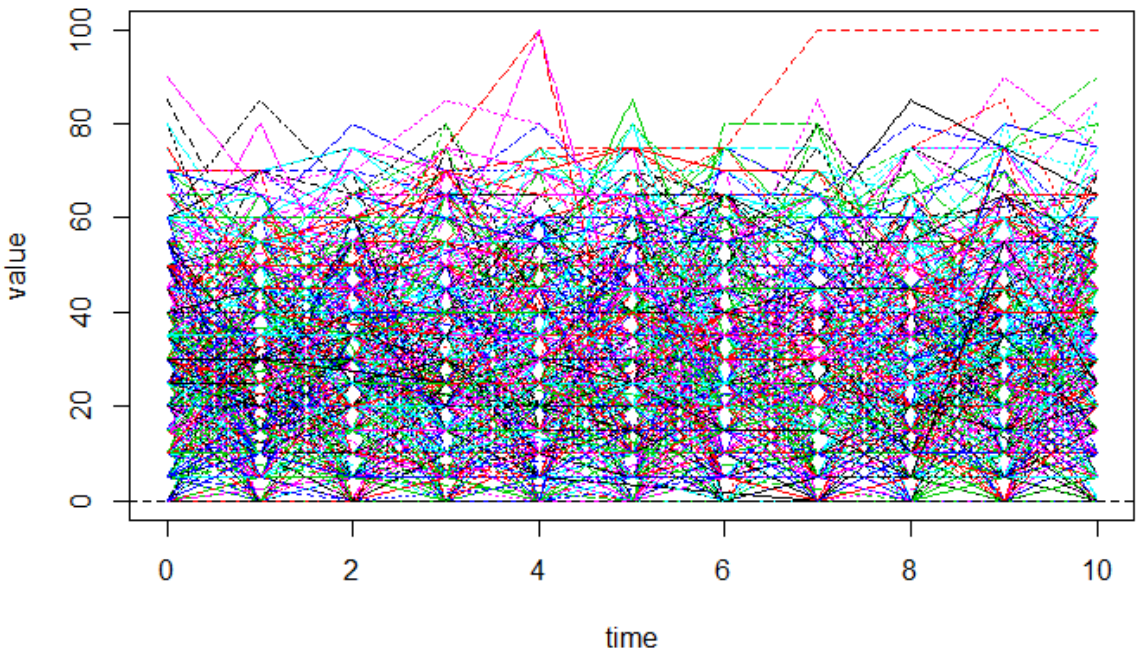


FIGURE 3.3: WOMAC Pain data reconstructed into their functional form.

3.4.C.) bases. We explored these three options along with the domain experts and concluded that a polygonal basis is the most suitable choice since Fourier series model periodic data (e.g., weather data following a full cycle each year) and b-splines are not convenient when we are trying to identify the peaks and valleys in the functional data since the curves are smoothed at the knots. Put another way, with a Fourier series basis our data would have to follow a complete cycle each year which is not the case for a slowly progressive disease. On the other hand, with a b-splines basis (see Figure 3.4.B.) represented by the seven functions of the third degree corresponding to three interior knots (placed at years 2, 5, and 8) shown as dotted vertical lines over the interval [0,10], are too smooth to capture the peaks. Using a polygonal basis allows us to capture the straight line trajectory between periods, without obscuring the peaks and valleys in the data. There are more options for basis systems but the creators of the `fda` package have not considered them common enough to include them in the code (Ramsay, Hooker, and Graves, 2009).

We specify the basis system (see Listing E.2) by using the `create.polygonal.basis` function from the `fda` package and pass the `argvals` argument with the location of the join points (e.g. at years 0, 2, 5, 8, and 10). `argvals` is defined as "a strictly increasing vector of argument values at which line segments join to form a polygonal line" (Ramsay et al., 2018, p. 55).

Running the `funHDDC` algorithm. The `funHDDC` algorithm is based on a functional latent mixture model that fits the data into group-specific functional subspaces (i.e., features and values best describing the group) via multivariate functional principal component analysis (MFPCA). The algorithm assumes the number of curves contained in each cluster can be described into a low dimensional functional latent subspace particular to each group, such that $d_k < \mathbb{R}, k = 1, ..., K$ where d are the intrinsic dimensions and K is the number of clusters. The goal is to cluster the observed multivariate functional curves $\mathbf{X}_1, ..., \mathbf{X}_n$ into K homogeneous groups. The model requires functional data objects for the `data` argument, the `model`(s) to be tested and the number of clusters `K`. The number of clusters K to test can be defined a priori or an estimation procedure can be summoned by defining a range for K as in Listing E.3. For a complete list of `funHDDC` arguments, please refer to Schmutz, Jacques, and Bouveyron (2019).

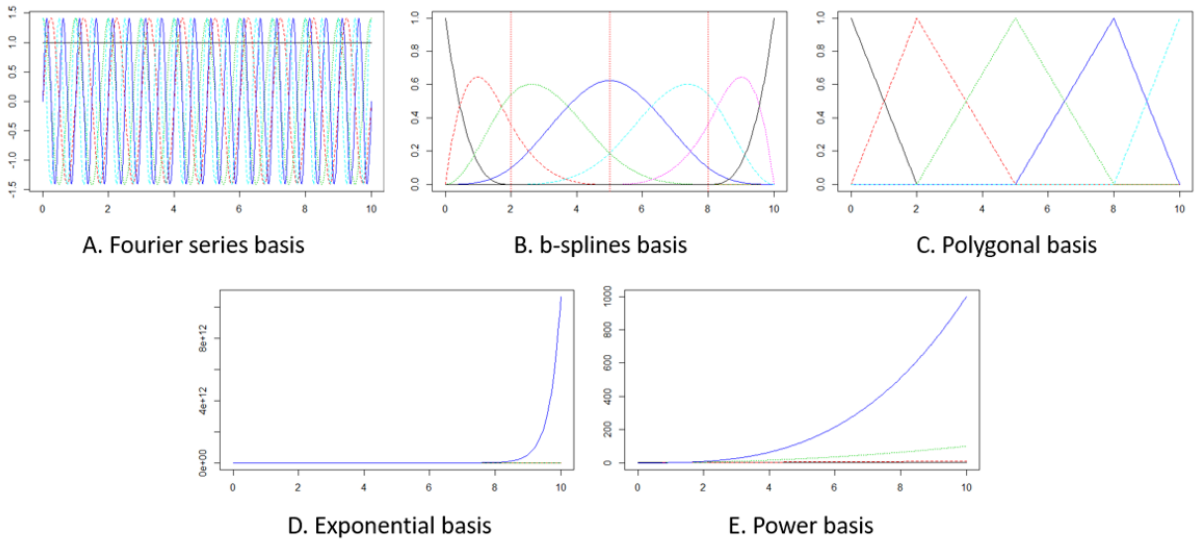


FIGURE 3.4: Illustration of five options of basis systems created with the `fda` package.

TABLE 3.11: Arguments passed to the funHDDC function, adapted from Schmutz, Jacques, and Bouveyron (2019).

Arguments	**Explanation**
data	In the multivariate case: a list of functional data objects.
K	The number of clusters (a single number, a set or a range)
model	The chosen model among 'AkjBkQkDk', 'AkjBQkDk', 'AkBkQkDk', 'ABkQkDk', 'AkBQkDk', 'ABQkDk'. 'AkjBkQkDk' is the default. Multiple models can be tested at the same time.

The package `funHDDC` proposes six different models, varying in levels of parsimony. The default model is named $a_{kj}b_kQ_kd_k$ from which five submodels are derived depending on the constraints applied on the parameters, resulting in more parsimonious models. Table 3.12 lists the definitions for each of the model parameters.

TABLE 3.12: funHDDC model parameters.

Parameter	**Definition**
a	Values of first diagonal elements of covariance matrix
k	Each of the clusters; instance of K
j	Number of observations
b	Values of last diagonal elements of covariance matrix
Q	Orthogonal matrix containing the basis expansion coefficients of the eigenfunctions
d	Number of dimensions for each cluster

Furthermore, `funHDDC` uses the EM (Expectation-Maximization) algorithm for inference of the model parameters. By default, 20 initializations of the EM algorithm are performed and the solution which maximizes the log-likelihood is presented. The expectation (E) step computes the conditional expectation of the complete log-likelihood using current parameters. The maximization (M) step maximizes the expected complete log-likelihood conditionally on the posterior probabilities estimated during the E step.

The final cluster result for an observation $x(t)$ is obtained by estimating its probability of belonging to each K cluster through the latent mixture model. Basically, `funHDDC` clusters multivariate functional data by projecting them into low dimensional subspaces. The projections are achieved by applying MFPCA to each cluster iteratively (Schmutz et al., 2020).

The code used to run the `funHDDC` algorithm can be seen in Listing E.3. The function prints the name of the models tested and the options chosen for the algorithm, the complexity of the model chosen (i.e., the number of free model parameters), and the Bayesian Information Criterion (BIC) score which can aid model selection. In this case, BIC is to be maximized since it is defined as $2LogLik - k\ln(n)$ where k is the number of parameters and n is the number of observations. Put differently, BIC is a negative number, thus we are searching for the maximum score (i.e., smallest negative number). In addition, the function prints the name of the best model according to the BIC criterion and the recommended number of clusters. An example of the output can be seen in Figure 3.5. Furthermore, several additional parameters can be extracted from the output, such as the proportion of individuals in each cluster (`prop`), the posterior probability for each individual belonging to each cluster (`posterior`), the clustering partition (`class`), the BIC scores, among others. For a full list of values obtained from running `funHDDC`, please see Schmutz, Jacques, and Bouveyron (2019).

```
funHDDC:

SELECTED: model AKJBQKDK with 3 clusters.
Selection Criterion: BIC.
```

	model <chr>	K <chr>	threshold <chr>	complexity <chr>	BIC <chr>
1	AKJBQKDK	3	0.2	69	-77,647.90
2	AKJBQKDK	4	0.2	102	-77,874.98
3	ABQKDK	3	0.2	76	-77,906.85
4	AKBQKDK	3	0.2	78	-77,917.59
5	AKJBQKDK	5	0.2	125	-77,940.08
6	AKBQKDK	4	0.2	110	-78,052.11
7	AKJBQKDK	6	0.2	148	-78,090.75
8	AKJBQKDK	7	0.2	171	-78,142.78
9	AKBQKDK	5	0.2	133	-78,229.12
10	ABQKDK	5	0.2	129	-78,289.07

FIGURE 3.5: Example of output of the `funHDDC` algorithm.

3.6.2 Setting Up Visualizations

This activity encompasses using the previously prepared data and creating descriptive visualizations. It is advised to start with simple visualizations and moving to more complex graphics later in the iterations. In order to show the results in an interpretable way, line chart plots were selected. Listing E.4 shows an example of how to create the graphical representation of cluster mean groups using the `plot` function from the *graphics* package (R Core Team, 2013). Figure 3.6 presents the resulting line chart which represents the five clusters' mean values for the variable WOMAC Pain (*wmpyns*).

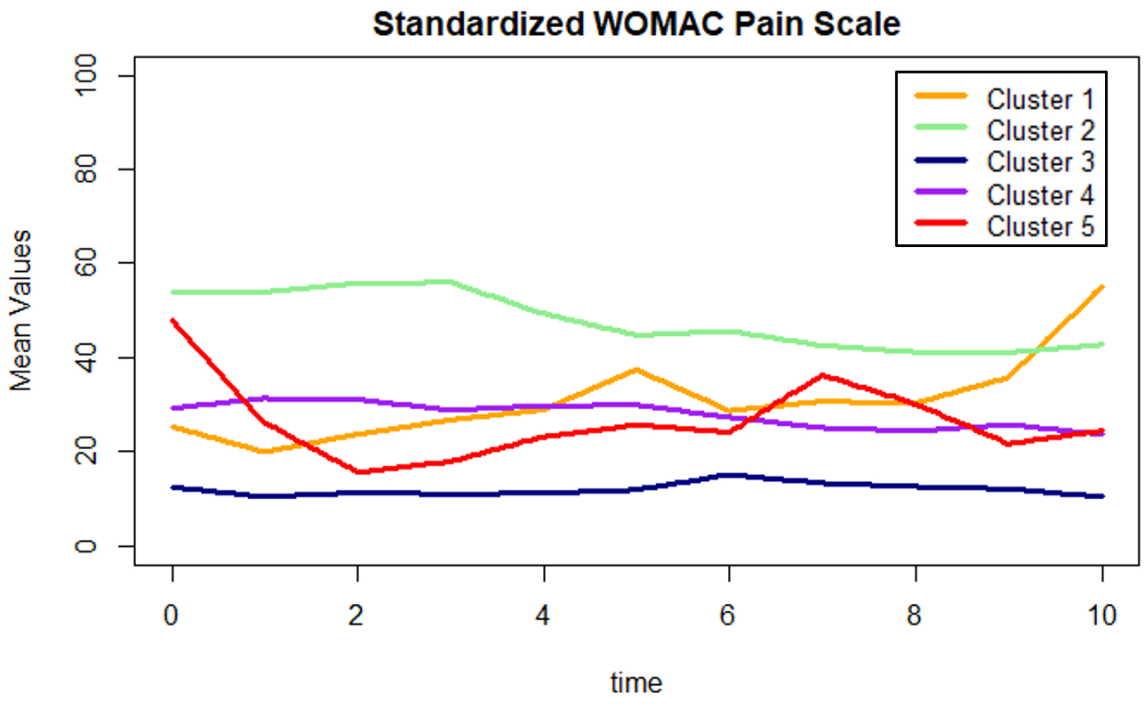


FIGURE 3.6: Example of visualization.

3.6.3 Exploring the Data

During this activity, discussions repeatedly take place among domain experts themselves and with the data scientist. The objective is to understand the information presented and decide on next steps of analysis, such as including or excluding features, different feature combinations, different machine learning algorithms, etc. In total, 28 iterations of data exploration occurred. During the iterations, mainly the selection and combination of features, use of basis systems, and number of clusters were explored. Appendix B shows the steps, outcomes, and participants of each iteration. Note that the data exploration iterations were mainly led by domain experts in terms of feature selection and clinical relevance. The data scientist typically lacks knowledge of the clinical domain but contributes by facilitating the understanding and application of the algorithms and the visualizations by writing the scripts and creating/presenting the plots. Nonetheless, team effort was needed to achieve the overall objectives and answer the research questions.

After consensus was reached for the last funHDDC clusters (during the last iteration), the comparison with the HCA exercise was performed.

Hierarchical Cluster Analysis. In order to create the HCA clusters, we used the `dist`, `hclust`, and `cutree` functions from the `stats` package (R Core Team, 2019). Listing E.5 shows the script used for the HCA analysis and Listing 3.2 shows the pseudocode. We executed the following steps:

1. We created a matrix with all the data points used in the funHDDC exercise. We used exactly the same data, which is already sampled (i.e., maximum of two data points missing) and with missing values imputed with linear interpolation.

2. We computed the distances with Euclidean method with the `dist` function and stored the resulting distance measures in an object of class `dist`.

3. We fed the distance matrix into the `hclust` function, and selected the agglomeration method within the `method` argument. The choices are "*ward.D2*", "*single*", "*complete*", "*average*" (= UPGMA), "*mcquitty*" (= WPGMA), "*median*" (= WPGMC) or "*centroid*" (= UPGMC). According to Hastie, Tibshirani, and Friedman (2009), given strong clustering tendency (as observed by Hopkins statistic in our case), since each of the clusters would be well-separated from the other, single, complete and average linkage methods yield similar results. The Ward method is the most commonly used in the literature for clustering knee OA phenotypes. Thus, we explored the Ward and average linkage methods.

4. We proceeded to use the `cutree` function to select the number of clusters we wished to inspect. We can plot the HCA dendrogram with the `plot` function and visually select the desired number of clusters by deciding on a particular dendrogram height which specifies the order in which the clusters were joined. However, in our case, we selected the same number of clusters as was decided during data exploration for the funHDDC clusters.

5. We plotted the group means for each feature set.

6. We visually inspected the graphical representation of cluster means for funHDDC and HCA clusters to compare for clinical relevance.

```
Prepare the data as a matrix
Compute Euclidean distance
Create dendrogram with desired linkage method
Plot dendrogram
Cut dendrogram tree at desired number of clusters K
Create list of members per cluster
Select sample with (previously created) observations index
Create dataframe with subject id and cluster membership
Create dataset with id and clusters included
Compute means of clusters
```

LISTING 3.2: Pseudocode for computing HCA clusters.

3.7 Inferential Analysis

Depending on the type of data at hand (e.g., continuous, discrete, binary, nominal, ordinal), we can perform different statistical tests to determine the statistical significance of the derived phenotypes. As stated in the Related Work chapter, if the data were normally distributed we could use the one-way ANOVA test. However, from the previously performed visual inspection of the data and Shapiro-Wilk test results (see Table 3.8), we determined the data is not normally distributed. Consequently, in order to compute the inter-cluster statistical significance, we used the Kruskal-Wallis rank sum test.

In addition, to estimate the agreement between cluster membership identified by funHDDC and HCA, we used the adjusted Rand index (ARI).

3.8 Deployment

The deployment phase commonly focuses on implementing the results discovered in the modeling and evaluation and inferential analysis phases. However, transforming the results into daily practice is a long process that falls outside the scope of this thesis. Therefore, we adapted this phase to encompass the communication of the analysis and results via a thesis report and a scientific article. Additionally, Figure 3.7 presents a Process-Deliverable Diagram (PDD) (Van De Weerd and Brinkkemper, 2009) which is a diagram that describes a process through the use of phases, activities, sub-activities and concepts. Concepts are deliverables resulting from the activities. Our PDD was created to detail our solution's phases, activities and deliverables and guide researchers through similar projects in the future.

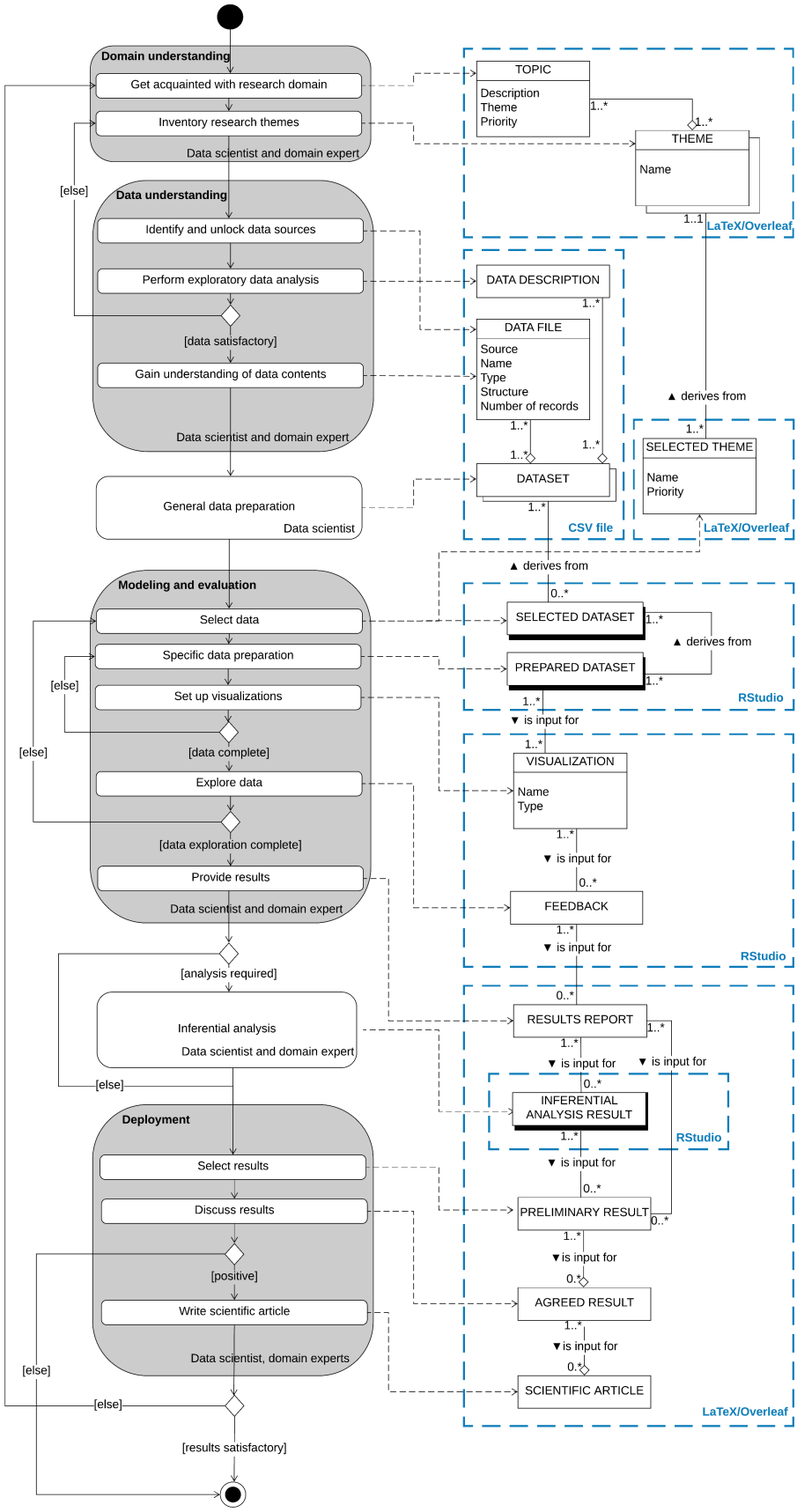


FIGURE 3.7: Process-Deliverable Diagram of solution using CRISP-IDM method.

Chapter 4

Results and Discussion

After conducting CRISP-IDM's phases beginning with domain understanding and ending with deployment, in this chapter we present our results and answer research questions **SQ2**, **SQ3**, and **MRQ**. Additionally, we used the consensus-based framework by Van Spil et al. (2020) as a *checklist* of the reporting recommendations for statistical analysis, specifically for the following: (i) availability of a pre-specified statistical analysis plan, (ii) analytical approach (supervised or unsupervised) and rationale, and (iii) criteria for clinical relevance and/or applicability and whether these were predefined.

4.1 Datasets: CHECK

After performing 28 iterations of data exploration, the domain experts estimated that clinically-relevant results were found. Five main groups were reached: WOMAC Pain, WOMAC Function, WOMAC Stiffness, KIDA, and OA Scoring. The first three groups are univariate and the latter two are multivariate scenarios.

The final set of features included in the analysis can be seen in Table 4.1. The number of records before sampling and imputation with linear interpolation were 1,002 for subjects and 2,004 for knees. Four variables for the KIDA analysis were created by computing the respective means of the femur and tibia measures: Lateral and Medial Osteophytes, and Lateral and Medial Bone Density.

Domain expertise was used to decide to cluster WOMAC measures at the subject level (not available at knee level) and radiographic measures (i.e., KIDA and OA Scoring) at the knee level. The rationale behind this decision is that joints within a patient may have a different underlying OA etiology (e.g., no OA in one joint and OA in another joint), which implicates having a different phenotype for each joint. In addition, as 'left' and 'right' knees are interchangeable regarding phenotypes in a patient (the designation is arbitrary in this context), it was deemed that clustering at joint level makes the most clinical sense.

TABLE 4.1: Final list of features included in the analysis.

Name	**Explanation**	**Scale**	**Years available**	**Records after sampling and imputation**	**Percentage data lost to sampling**
Questionnaire data: WOMAC - at subject level					
wmpyns	WOMAC Pain: standardized pain scale	0-100	All (0-10)	819	18%
wmfuns	WOMAC Function: standardized physical functioning scale	0-100	All (0-10)	819	18%

Table 4.1 continued from previous page

Name	**Explanation**	**Scale**	**Years available**	**Records after sampling and imputation**	**Percentage data lost to sampling**
wmstfs	WOMAC Stiffness: standardized stiffness scale	0-100	All (0-10)	820	18%
Radiographic variables: Knee Images Digital Analysis (KIDA) - at knee level					
OsteophyteLatmm	Lateral Osteophytes: mean of femur and area	R≥0 in mm^2	0,2,5,8,10	1888	6%
OsteophyteMedmm	Medial Osteophytes: mean of femur and tibia area	R≥0 in mm^2	0,2,5,8,10	1888	6%
MeanLatJSWmm	Lateral Joint Space Width (mean)	R≥0 in mm	0,2,5,8,10	1889	6%
MeanMedJSWmm	Medial Joing Space Width (mean)	R≥0 in mm	0,2,5,8,10	1888	6%
BDMeanLatmmAl	Lateral Bone Density: mean of femur and tibia area	R≥0 in mmAl	0,2,5,8,10	1846	8%
BDMeanMedmmAl	Medial Bone Density: mean of femur and tibia area	R≥0 in mmAl	0,2,5,8,10	1846	8%
			Combined:	1788	11%
Radiographic variables: OA Scoring (skyline views) - at knee level					
K_SKY_SCL	Knee sky patellofemoral sclerosis	0,1,2,3	0,2,5,8,10	1788	11%
K_SKY_JSN	Knee sky patellofemoral narrowing	0,1,2,3	0,2,5,8,10	1877	6%
K_SKY_OST	Knee sky patellofemoral osteophytes	0,1,2,3	0,2,5,8,10	1873	7%
			Combined:	1788	11%

4.2 Analysis

In order to understand how well a MBCFD method performs at identifying clinically-relevant and statistically-significant phenotypes, we selected funHDDC based on its ability to handle multivariate functional data, interpretability, and flexibility in modeling the data since it uses *adaptive* basis expansion systems. Jacques and Preda (2014b) proposed the first model-based algorithm for functional data which has the advantage of incorporating the dependence among curves into the analysis. However, it was not until Schmutz et al. (2020) that this method was expanded to the multivariate case by using a functional latent mixture model with MFPCA. In addition, we selected HCA as the method to compare with funHDDC since, according to the answer of **SQ1**, 24% of the surveyed literature that aimed to find knee OA phenotypes used HCA as their primary method.

In order to evaluate the clinical relevance, we utilized the previously described list of four criteria: (i) different inter-cluster feature trajectories, (ii) upward/downward trajectories, (iii)

multi-feature different behavior, (iv) balanced number of clusters. For the univariate case, only the second and fourth criteria apply.

For the evaluation of statistical significance, we chose the Kruskal-Wallis rank sum test to determine if there are statistically-significant differences between the derived clusters for the 21 baseline characteristics listed in Table 4.2.

TABLE 4.2: Meaning of baseline characteristics.

	Variable	**Meaning**
1	Lft_T0	Age, mean± sd
2	RAS	Race, white %
3	SEXE	Sex, female %
4	BMI	Body Mass Index (kg/m^2), mean±sd
5	Menopauze_01	Menopause, no. (% post)
6	Leptinengml	Leptine (ng/ml), mean±sd
7	Adiponectineugml	Adiponectine (ug/ml), mean±sd
8	Resistinengml	Resistine (ng/ml), mean±sd
9	CTXIugmmol	CTX-I (ug/mmol) (C-terminal telopeptide of collagen I), mean±sd
10	uNTXInMBCEmmol	N-terminal telopeptide of collagen I (nM BCE/mmol), mean±sd
11	sPINP	Aminoterminal propeptide of type I procollagen, mean±sd
12	sOC	Osteocalcin mean±sd
13	sC12C	sC1,2C (collagen of type I and II), mean±sd
14	CTXIIngmmol	CTX-II (ng/mmol) (C-terminal telopeptide of type II collagen), mean±sd
15	sCS846	Chondroitin sulphate 846, mean±sd
16	sCOMPUl	sCOMP (μg/ml) (cartilage oligomeric matrix protein), mean±sd
17	sPIIANP	Collagen N-propeptide of type IIA, mean±sd
18	sHA	Hyaluronic acid, mean±sd
19	sPIIINP	N-terminal propeptide of type III procollagen, mean±sd
20	hsCRP	High-sensitivity C-reactive protein, mean±sd
21	BSE	Erythrocyte sedimentation rate, mean±sd

Lastly, one way to compare unsupervised clustering models is to apply them on data for which we know the solution (i.e., with ground truth labels) so that we are able to estimate the percentage of error for each model and select the best, based on which model made less errors. However, we do not have access to ground truth CHECK data. Put differently, if we had a *gold standard dataset*, we could compute metrics such as RMSE (root mean squared error) which tells us the difference between the values of the predictions and the estimates. Moreover, metrics like the Silhouette index (see section 1.3.1 which computes intra- and inter-cluster distances is only useful to compare distance-based models. Formally, a definition for distance in functional statistics does not exist. In addition, we could compute the observed distances between the mean curves of the clusters by discretizing the data, however, it is difficult to determine which clusters to compare because the associations between the groups in both clustering exercises are not always clear.

Thus, we used the adjusted Rand index (ARI) to evaluate agreement between the funHDDC and HCA clustering exercises. We used the `ari` function from the `CrossClustering` package which computed the ARI and the confidence interval.

4.2.1 WOMAC Pain Phenotypes

The WOMAC questionnaire features are available at the subject level (i.e., no measurements were taken at joint level during data collection) were initially considered to be clustered together, however, we did not find any combination of features or number of clusters where the trajectories were different between features and clusters. For example, if WOMAC Pain was increasing, so was WOMAC Function, and this finding is not deemed as clinically relevant as it contradicts clinical intuition and previous findings. Following, when we clustered these features separately, we were able to find interesting trajectories, thus it was decided to keep WOMAC features separately in the analysis.

Clinical Relevance

The domain expertise decision for WOMAC Pain clusters was to evaluate five phenotypes that agreed with the funHDDC's selection of best number of clusters by BIC score. The WOMAC Pain clusters derived were named as: (i) increasing pain, (ii) constant high pain, (iii) constant low pain, (iv) constant moderate pain, and (v) initial high then moderate pain. Figure 4.1 shows the final five funHDDC clusters in comparison with HCA clusters using the Ward and average linkage methods.

The funHDDC clusters complied with the two clinical relevance criteria for the univariate case: (ii) upward/downward trajectories (cluster one and five), and (iv) balanced number of clusters as funHDDC was able to find progression patterns in the K=5 solution, whereas HCA shows some progression at K=5 and K=8 but not as marked as funHDDC. Five clusters were deemed as a relevant number of clusters by domain experts.

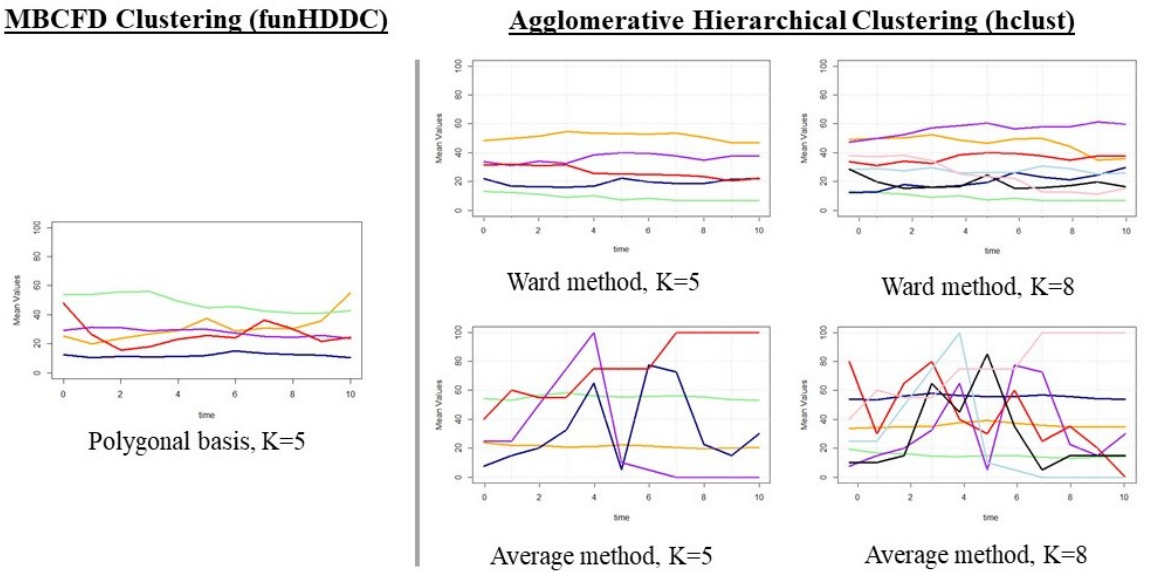
The HCA clusters with Ward method did not comply with either of the two clinical relevance criteria, as no meaningful upward/downward trajectories were found in the K=5 solution. The K=8 solution presents a downward trajectory for cluster eight, however, the K=8 solution does not comply with the balanced number of clusters criterion.

For the HCA exercise with average linkage method, clusters are discarded since the trajectories do not make clinical sense. Moreover, OA is a slowly progressive disease that does not show such erratic behavior. In general, the average linkage method is frequently used (Hastie, Tibshirani, and Friedman, 2009), however, this method might not be appropriate for clustering longitudinal data since it creates round-shaped clusters due to using the average distance between all pairs of observations. Additionally, cluster membership for the K=5 scenario was highly imbalanced with cluster one having 92% of all subjects. For funHDDC results, the cluster with the most members is cluster three with 39% of all subjects.

Statistical Significance

Descriptive statistics for each of the five clusters can be found in Appendix C, Table C.1 for funHDDC clusters and Table C.6 for HCA clusters. We evaluated the 21 baseline characteristics listed in Table 4.2.

Age (*Lft_T0*) and collagen type I and II (*sC12C*) were found to be statistically significant ($p \leq 0.05$). The average posterior probabilities for the funHDDC clusters were 0.94 for 819 members. Table D.1 shows the posterior probabilities for funHDDC clusters in detail.



Cluster membership MBCFD

Cluster	1	2	3	4	5	Total
Subjects	66	77	315	312	48	819
Percentage	8%	9%	39%	38%	6%	100%

Cluster membership HCA (Ward method)

Cluster	1	2	3	4	5	Total
Subjects	92	248	200	136	143	819
Percentage	11%	30%	24%	17%	17%	100%

FIGURE 4.1: WOMAC Pain clusters for MBCFD and HCA.

Clustering Validation

In order to measure the similarity between the funHDDC and HCA clustering exercises, we computed the ARI. The resulting ARI was < 0.65 indicating poor recovery, with confidence interval [0,0].

4.2.2 WOMAC Function Phenotypes

Clinical Relevance

The domain expertise decision for WOMAC Function clusters was to evaluate five phenotypes that agreed with the funHDDC's selection of best number of clusters by BIC score. The WOMAC Function clusters derived were named as: (i) constant high functional limitation, (ii) constant low functional limitation, (iii) increasing functional limitation, (iv) constant moderate functional limitation, and (v) fluctuating moderate functional limitation. Figure 4.2 shows the final five funHDDC clusters in comparison with HCA clusters using the Ward and average linkage methods.

The funHDDC clusters complied with the two clinical relevance criteria for the univariate case: (ii) upward/downward trajectories (cluster three and five), and (iv) balanced number of clusters as funHDDC was able to find progression patterns in the K=5 solution, whereas HCA shows some progression at K=5 and K=8 but not as marked as funHDDC.

The HCA clusters with Ward method partially complied with the two clinical relevance criteria, as a downward trajectory was found in the K=5 solution, but no upward trajectory was detected. The K=8 solution presents an upward trajectory for cluster five but not as pronounced

as funHDDC. Moreover, the K=8 solution does not comply with the balanced number of clusters criterion as the clinically-relevant number was decided at five.

For the HCA exercise with average linkage method, cluster membership for the K=5 scenario was highly imbalanced with cluster two having 90% of all subjects. For funHDDC results, the cluster with the most members is cluster four with 39% of all subjects.

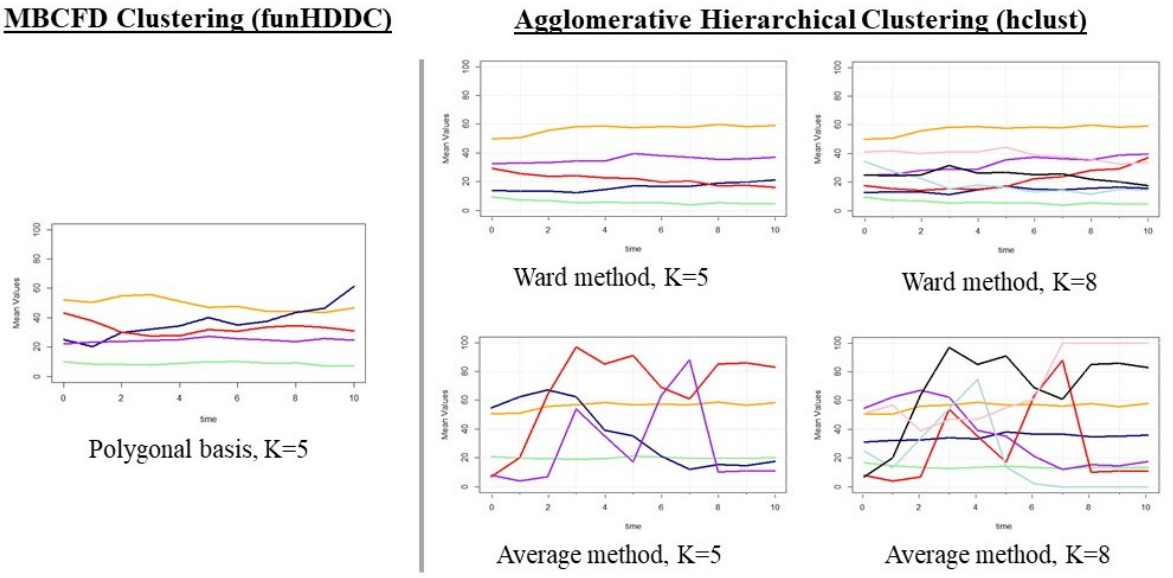
Statistical Significance

Descriptive statistics for each of the five clusters can be found in Appendix C, Table C.2 for funHDDC clusters and Table C.7 for HCA clusters. We evaluated the 21 baseline characteristics listed in Table 4.2.

Collagen N-propeptide of type IIA (*sPIIANP*) was found to be statistically significant ($p \leq 0.05$). The average posterior probabilities for the funHDDC clusters were 0.94 for 819 members. Table D.1 shows the posterior probabilities for funHDDC clusters in detail.

Clustering Validation

In order to measure the similarity between the funHDDC and HCA clustering exercises, we computed the ARI. The resulting ARI was < 0.65 indicating poor recovery, with confidence interval [0,0].



Cluster membership MBCFD

Cluster	1	2	3	4	5	Total
Subjects	80	274	37	323	105	819
Percentage	10%	33%	5%	39%	13%	100%

Cluster membership HCA (Ward method)

Cluster	1	2	3	4	5	Total
Subjects	69	186	188	202	174	819
Percentage	8%	23%	23%	25%	21%	100%

FIGURE 4.2: WOMAC Function clusters for MBCFD and HCA.

4.2.3 Overlap between WOMAC Pain and WOMAC Function

We can observe in Figure 4.3 that WOMAC Pain and WOMAC Function funHDDC clusters do not overlap. This was the reason to cluster them separately even though the initial plan was to cluster Pain and Function together. These results were deemed very interesting by domain experts and further research will be pursued to investigate these phenomena. For example, the stacked bar chart shows that subjects with increasing pain do not necessarily show increasing problems with functional limitation, which would be the clinical expectation.

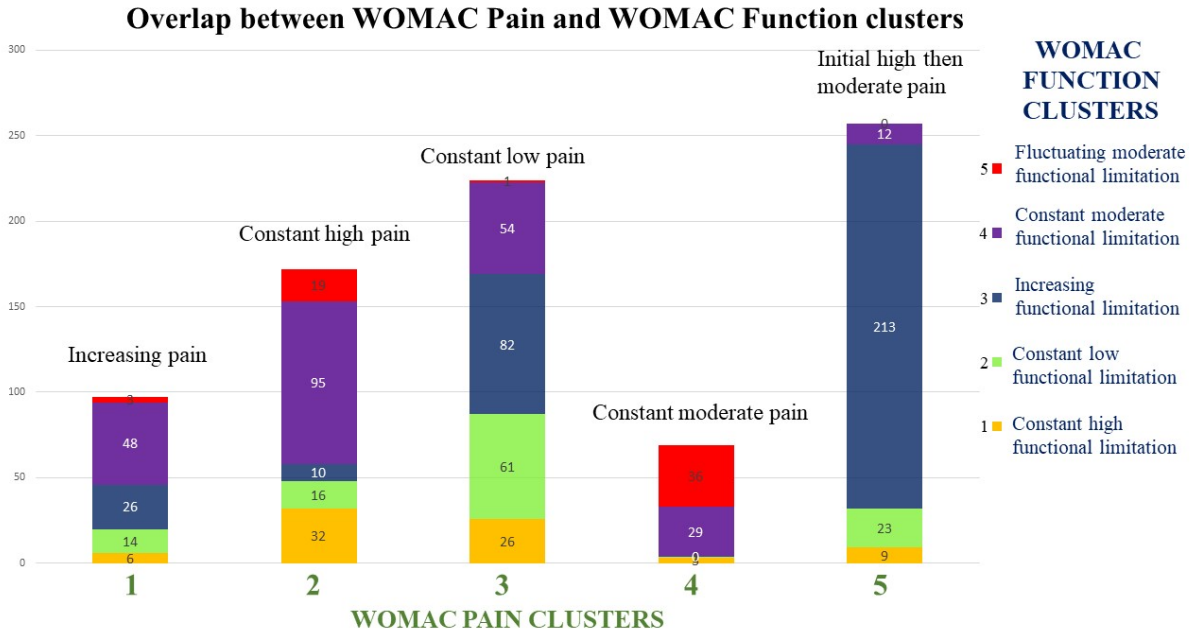


FIGURE 4.3: Overlap between WOMAC Pain and WOMAC Function clusters for MBCFD analysis.

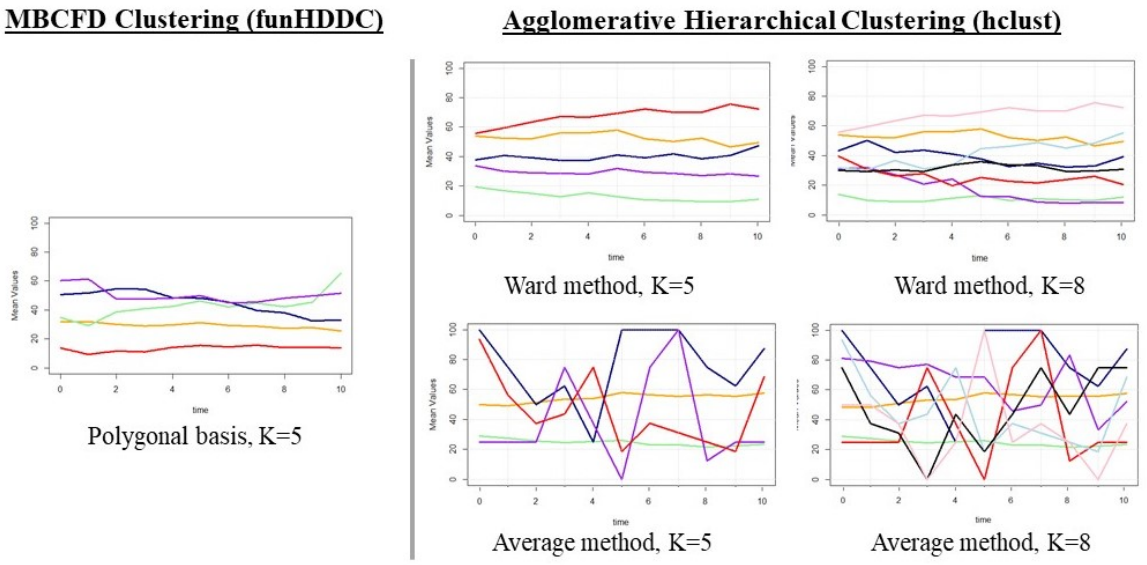
4.2.4 WOMAC Stiffness Phenotypes

Clinical Relevance

The domain expertise decision for WOMAC Stiffness clusters was to evaluate five phenotypes that agreed with the funHDDC's selection of best number of clusters by BIC score. The WOMAC Stiffness clusters derived can be named as: (i) constant moderate stiffness, (ii) fluctuating then high stiffness, (iii) decreasing stiffness, (iv) high then moderate stiffness, (v) constant low stiffness. Figure 4.4 shows the final five funHDDC clusters in comparison with HCA clusters using the Ward and average linkage methods.

The funHDDC clusters complied with the two clinical relevance criteria for the univariate case: (ii) upward/downward trajectories (cluster two, three and four), and (iv) balanced number of clusters as funHDDC was able to find significant progression patterns in the K=5 solution, whereas HCA shows significant progression at K=8.

The HCA clusters with Ward method did not comply with either of the two clinical relevance criteria, as no meaningful upward/downward trajectories were found in the K=5 solution. The K=8 solution presents a downward trajectory for cluster four and an upward trajectory for cluster six, however, the K=8 solution does not comply with the balanced number of clusters



Cluster membership MBCFD

Cluster	1	2	3	4	5	Total
Subjects	325	91	89	126	189	820
Percentage	40%	11%	11%	15%	23%	100%

Cluster membership HCA (Ward method)

Cluster	1	2	3	4	5	Total
Subjects	113	240	164	256	47	820
Percentage	14%	29%	20%	31%	6%	100%

FIGURE 4.4: WOMAC Stiffness clusters for MBCFD and HCA.

criterion as funHDDC presents upward/downward trajectories with fewer number of clusters (K=5).

For the HCA exercise with average linkage method, cluster membership for the K=5 scenario was highly imbalanced with cluster two having 75% of all subjects. For funHDDC results, the cluster with the most members is cluster four with 40% of all subjects.

Statistical Significance

Descriptive statistics for each of the five clusters can be found in Appendix C, Table C.3 for funHDDC clusters and Table C.8 for HCA clusters. We evaluated the 21 baseline characteristics listed in Table 4.2.

Aminoterminal propeptide of type I procollagen (*sPINP*) and erythrocyte sedimentation rate (*BSE*) were found to be statistically significant ($p \leq 0.05$). The average posterior probabilities for the funHDDC clusters were 0.89 for 820 members. Table D.1 shows the posterior probabilities for funHDDC clusters in detail.

Clustering Validation

In order to measure the similarity between the funHDDC and HCA clustering exercises, we computed the ARI. The resulting ARI was < 0.65 indicating poor recovery, with confidence interval [0,0].

4.2.5 KIDA Phenotypes

Clinical Relevance

The domain expertise decision for KIDA clusters was to evaluate eight phenotypes that agreed with the funHDDC's selection of best number of clusters by BIC score. The domain expertise expectation was to find distinction between lateral and medial phenotypes, however, we only found this distinction on osteophytes. All clusters had increasing lateral joint space width (JSW) and decreasing medial JSW. The KIDA clusters derived can be named as: (i) Low osteophytes, moderate bone density (ii) increasing bone density, (iii) low bone density, (iv) increasing lateral osteophytes, low medial osteophytes, moderate bone density, (v) increasing lateral and medial osteophytes, low medial JSW, high lateral JSW, increasing bone density; (vi) moderate bone density, (vii) high bone density, (viii) slightly increasing bone density. Figures 4.5 and 4.6 show the final eight funHDDC clusters in comparison with HCA clusters using the Ward and average linkage methods.

The funHDDC clusters complied with the four clinical relevance criteria for the multivariate case: (i) different inter-cluster feature trajectories (lateral and medial osteophytes), (ii) upward/-downward trajectories (present in all features), (iii) multi-feature different behavior (lateral and medial osteophytes). Even though eight clusters could be considered a high number, during the evaluation with domain experts, eight clusters were deemed acceptable. From our literature review, we found that six is the highest reported number of knee OA phenotypes. However, this is still under domain expertise discussion.

The HCA clusters complied with the four clinical relevance criteria for the multivariate case: (i) different inter-cluster feature trajectories (lateral and medial osteophytes, lateral and medial JSW), (ii) upward/downward trajectories (present in all features), (iii) multi-feature different behavior (lateral and medial osteophytes), and (iv) balanced number of clusters. Eight clusters were deemed as a relevant number of clusters by domain experts.

It is quite interesting to find that HCA (unlike funHDDC) was able to detect a constant if not slightly decreasing trajectory for lateral JSW, however, the caveat being that the eight cluster is rather small (n=18 knees). Thus, this insight cannot be yet be reported as significant. In this exercise, HCA found interesting trajectories, hence, it cannot be confirmed that funHDDC outperforms HCA in clinical relevance.

For the HCA exercise with average linkage method, clusters were discarded since the trajectories do not make clinical sense. For the HCA exercise with average linkage method, cluster membership for the K=8 scenario was highly imbalanced with cluster one having 98% of all subjects. For funHDDC results, the cluster with the most members is cluster six with 23% of all subjects.

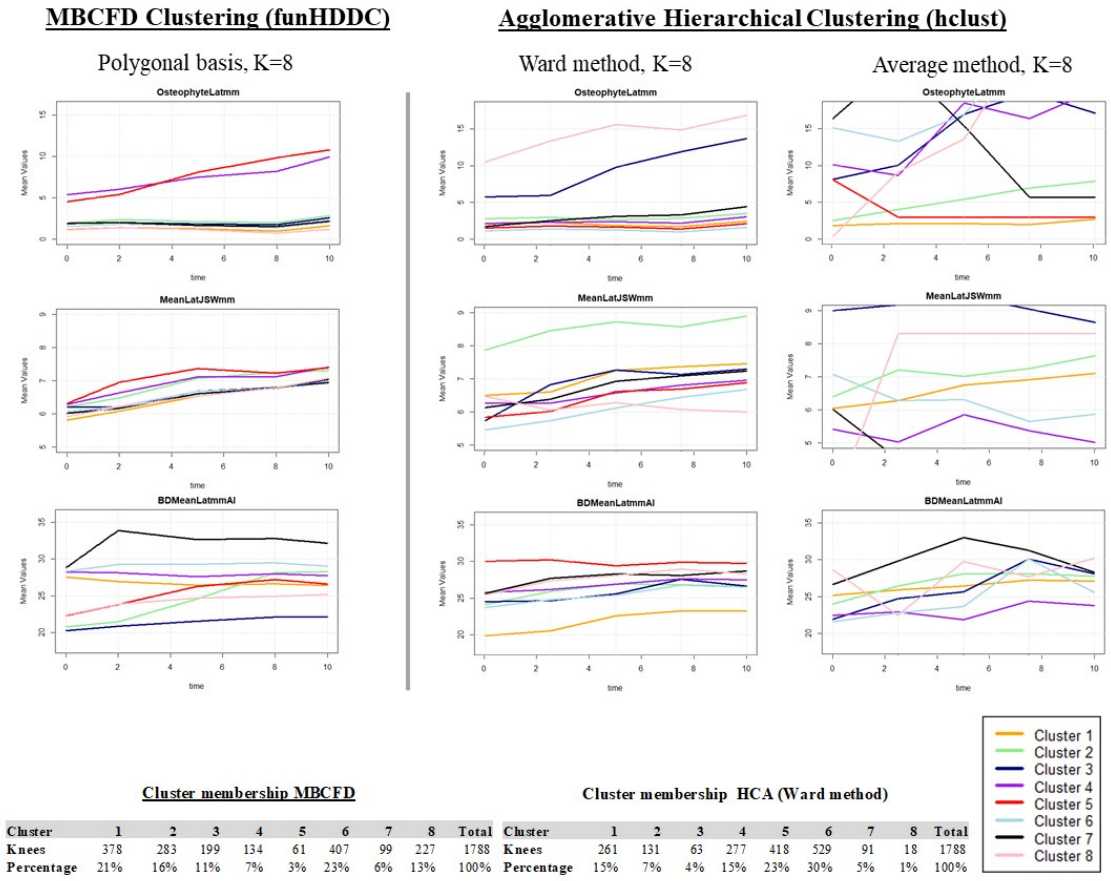
Statistical Significance

Descriptive statistics for each of the five clusters can be found in Appendix C, Table C.4 for funHDDC clusters and Table C.9 for HCA clusters. We evaluated the 21 baseline characteristics listed in Table 4.2 and found no statistically-significant differences between the groups. The average posterior probabilities for the funHDDC clusters were 0.96 for 1788 knees. Table D.1 shows the posterior probabilities for funHDDC clusters in detail.

Clustering Validation

In order to measure the similarity between the funHDDC and HCA clustering exercises, we computed the ARI. The resulting ARI was < 0.65 indicating poor recovery, with confidence

interval [0,0].



Cluster membership MBCFD

Cluster	1	2	3	4	5	6	7	8	Total
Knees	378	283	199	134	61	407	99	227	1788
Percentage	21%	16%	11%	7%	3%	23%	6%	13%	100%

Cluster membership HCA (Ward method)

Cluster	1	2	3	4	5	6	7	8	Total
Knees	261	131	63	277	418	529	91	18	1788
Percentage	15%	7%	4%	15%	23%	30%	5%	1%	100%

FIGURE 4.5: KIDA Lateral clusters for MBCFD and HCA.

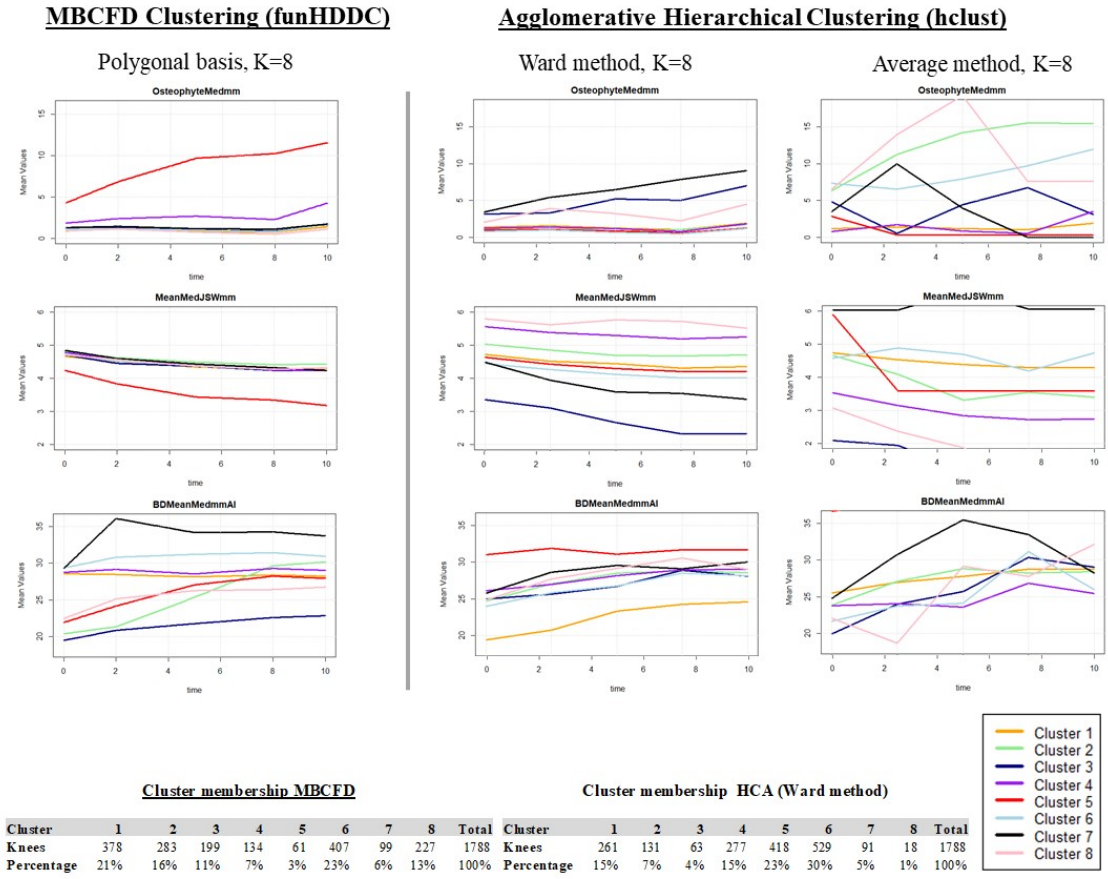
4.2.6 OA Scoring Phenotypes

Clinical Relevance

The domain expertise decision for OA Scoring clusters was to evaluate six phenotypes that agreed with the funHDDC's selection of best number of clusters by BIC score. The OA Scoring clusters derived can be named as: (i) Moderate osteophytes, (ii) low sclerosis, (iii) all low features, (iv) increasing osteophytes, (v) highly increasing sclerosis, increasing JSN, high osteophytes, (vi) high osteophytes. Figure 4.7 shows the final six funHDDC clusters in comparison with HCA clusters using the Ward linkage method for K=6 and K=8.

The funHDDC clusters complied with the four clinical relevance criteria for the multivariate case: (i) different inter-cluster feature trajectories (cluster four), (ii) upward/downward trajectories (clusters two, four and five), (iii) multi-feature different behavior (cluster four), and (iv) balanced number of clusters as six phenotypes have been previously reported in the literature and six clusters were deemed relevant by domain experts.

The HCA clusters complied with the four clinical relevance criteria for the multivariate case: (i) different inter-cluster feature trajectories (all clusters except cluster four), (ii) upward/-downward trajectories, (iii) multi-feature different behavior (all clusters except four), and (iv) balanced number of clusters as six phenotypes have been previously reported in the literature



Cluster membership MBCFD

Cluster	1	2	3	4	5	6	7	8	Total
Knees	378	283	199	134	61	407	99	227	1788
Percentage	21%	16%	11%	7%	3%	23%	6%	13%	100%

Cluster membership HCA (Ward method)

Cluster	1	2	3	4	5	6	7	8	Total
Knees	261	131	63	277	418	529	91	18	1788
Percentage	15%	7%	4%	15%	23%	30%	5%	1%	100%

FIGURE 4.6: KIDA Medial clusters for MBCFD and HCA.

and six clusters were deemed relevant by domain experts. It is interesting to find that HCA was able to detect significant progression for two of the three features. Thus, it cannot be confirmed that funHDDC outperforms HCA in clinical relevance for the OA Scoring exercise.

For the HCA exercise with average linkage method, clusters were discarded since the trajectories do not make clinical sense (not shown in Figure 4.7). For the HCA exercise with average linkage method, cluster membership for the K=6 scenario was highly imbalanced with cluster one having 93% of all subjects. For funHDDC results, the cluster with the most members is cluster one with 29% of all subjects.

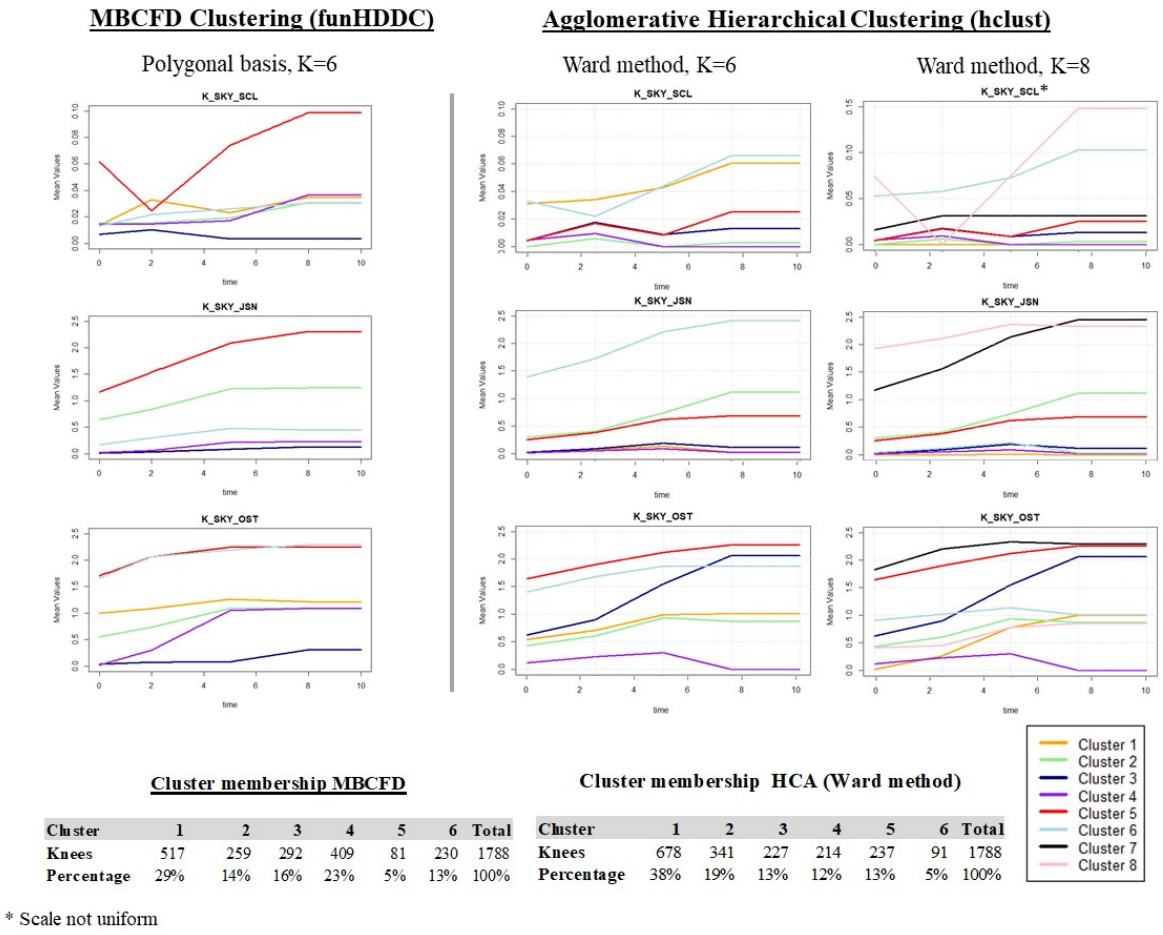
Statistical Significance

Descriptive statistics for each of the six clusters can be found in Appendix C, Table C.5 for funHDDC clusters and Table C.10 for HCA clusters. We evaluated the 21 baseline characteristics listed in Table 4.2 and found no statistically-significant differences between the groups. The average posterior probabilities for the funHDDC clusters were 0.99 for 1788 knees. Table D.1 shows the posterior probabilities for funHDDC clusters in detail.

Clustering Validation

In order to measure the similarity between the funHDDC and HCA clustering exercises, we computed the ARI. The resulting ARI was < 0.65 indicating poor recovery, with confidence

interval [0,0].



Cluster membership MBCFD

Cluster	1	2	3	4	5	6	Total
Knees	517	259	292	409	81	230	1788
Percentage	29%	14%	16%	23%	5%	13%	100%

Cluster membership HCA (Ward method)

Cluster	1	2	3	4	5	6	Total
Knees	678	341	227	214	237	91	1788
Percentage	38%	19%	13%	12%	13%	5%	100%

* Scale not uniform

FIGURE 4.7: OA Scoring clusters for MBCFD and HCA.

4.3 Answers to Research Questions SQ2, SQ3 and MRQ

Overall, we can state that funHDDC outperformed HCA in terms of clinical relevance and statistical significance for the univariate case, but not for the multivariate case. Table 4.3 shows a summary of these results per variable group. In the following sections, we provide an answer to the remaining research questions: **SQ2**, **SQ3** and **MRQ**.

4.3.1 SQ2: How well does the selected MBCFD method perform at identifying clinically-relevant and statistically-significant knee OA phenotypes?

In order to answer this question, we look at the top half of Table 4.3. funHDDC complies with the clinical relevance criteria for our five variable groups. However, in statistical significance, it only performs well for the univariate case (i.e., two, one and two statistically-significant baseline characteristics for WOMAC Pain, Function and Stiffness, respectively) as no differences between the groups were found for the multivariate case. As a reminder to the reader, the input

of domain experts was used to determine the clinical relevance of the subgroups throughout the interactive modeling and evaluation iterations' unstructured interviews.

TABLE 4.3: Results per variable group with regards to clinical relevance and statistical significance.

Group	**Clinical Relevance Criteria**				**Statistically-Significant Characteristics**
funHDDC Clusters					
Univariate	***(ii)***		***(iv)***		
WOMAC Pain	✓		✓		Age (Lft_T0); collagen of type I and type II (sC12C)
WOMAC Function	✓		✓		Collagen N-propeptide of type IIA (sPIIANP)
WOMAC Stiffness	✓		✓		Aminoterminal propeptide of type I procollagen (SPINP); Erythrocyte sedimentation rate (BSE)
Multivariate	***(i)***	***(ii)***	***(iii)***	***(iv)***	
KIDA	✓	✓	✓	✓	None found
OA Scoring	✓	✓	✓	✓	None found
HCA Clusters					
Univariate	***(ii)***		***(iv)***		
WOMAC Pain	x		x		None found
WOMAC Function	p		p		None found
WOMAC Stiffness	x		x		None found
Multivariate	***(i)***	***(ii)***	***(iii)***	***(iv)***	
KIDA	✓	✓	✓	✓	None found
OA Scoring	✓	✓	✓	✓	None found

The clinical relevance criteria are: (i) different inter-cluster feature trajectories, (ii) upward/ downward trajectories, (iii) multi-feature different behavior, (iv) balanced number of clusters. The checkmark indicates compliance, the x mark indicates non-compliance and p indicates partial compliance.

4.3.2 SQ3: How does the MBCFD method perform compared to a non-functional clustering method?

The selected MBCFD algorithm, funHDDC, outperforms HCA in clinical relevance in the univariate case due to its ability to detect upward/downward trajectories not picked up by HCA at all or only at a higher number of clusters. In terms of statistical significance, funHDDC's clusters presented two, one and two baseline characteristics that showed a significant difference, respectively. On the other hand, HCA's clusters did not show statically-significant differences.

Regarding the multivariate case, funHDDC performed similarly to HCA for clinical relevance and statistical significance. However, domain experts chose to pursue the funHDDC clusters.

One of the reasons funHDDC might have outperformed HCA is that it uses MFPCA separately per cluster, thus it is better equipped to model time-dependent trajectories that can be represented by the eigenfunctions as it assumes that data live in subspaces of different dimensions. Another advantage is funHDDC's modeling flexibility as it tries different combinations of model parameters and compares them by BIC score. On the other hand, HCA only has two parameters that can be modified: linkage method and distance method. Even though we testes two linkage methods (i.e., Ward and average), we only experimented with Euclidean distance.

One drawback of HCA is that it tends to form spherical clusters, which is not ideal when trying to model trajectories.

funHDDC is a clustering method that relies on a probabilistic model, thus contrary to some other methods it does not follow any assumption on data normality. Therefore, we can use this method on both normal and non-normal data. The only "strong" hypothesis of funHDDC is that the dataset contains independent individuals.

4.3.3 MRQ: To what extent can model-based clustering for functional data contribute to derive clinically-relevant and statistically-significant knee OA phenotypes?

In general, the extent to where MBCFD can contribute to derive knee OA phenotypes is full for the univariate case, but it is limited to clinical relevance in the multivariate case. On the one hand, for the univariate case, MBCFD yields better performance in deriving clinically-relevant and statistically-significant knee OA phenotypes than HCA. Concerning clinical relevance, MBCFD complied with the criteria by detecting upward/downward trajectories not identified by HCA whatsoever or only at a higher number of clusters. Moreover, MBCFD was able to detect five statistically-significant characteristics between the phenotypes whereas HCA did not detect any statistically-significant differences between the groups. On the other hand, for the multivariate case in both clustering exercises, the phenotypes were clinically relevant by complying with the criteria, but no statistical significance was found between the groups.

4.4 Threats to Validity

In this section, we discuss identified threats to the validity of this experiment, including mitigation strategies. The threats are presented according to the four different types of validity based on Wohlin et al. (2012).

Construct validity refers to the understanding of the constructs included in the research. The main threat for us is the misinterpretation of concepts related to functional data analysis, particularly model-based clustering of functional data (MBCFD), and OA, such as OA heterogeneity, phenotypes and constructs specific to the domain. The main risk lies in *combining* these two fields such that the results from the study remain valid. To mitigate these threats, we involved experts from both domains. In addition, when performing the semi-systematic literature review, we mimicked the search strategy from experienced researchers in knee OA phenotypes. We also followed three of the reporting recommendations for statistical analysis from the consensus-based framework for conducting and reporting osteoarthritis phenotype research by Van Spil et al. (2020).

Another threat in this category is related to whether tests used measure what they are intended to measure. Unfortunately, in unsupervised learning, it is not possible to measure accuracy nor is it designed to predict future phenotype membership. The results are informative but not definitive. Thus, we used statistical tests such as Hopkins statistic to determine whether we had potentially meaningful insights in the data, Kruskal-Wallis rank sum test to see if there were statistically-significant differences between the groups, and ARI metric to check the overlap between the two clustering exercises.

Internal validity focuses on how systematic error is minimized and how the experiments were conducted. In terms of our literature search, the main threat is potential bias and subjective interpretation when examining available literature by the author. To mitigate this issue we

began with studies performed with the same data and purpose, performed a semi-systematic literature search following an expert-designed search strategy, and received the guidance of domain experts. To increase coverage, we used forward and backward *snowballing* to identify additional papers. However, the inclusion of additional databases could have yielded complementary insights. With regards to our experiments, different experts were involved throughout the data exploration steps, each with their own expertise and/or clinical intuition. In order to mitigate potential bias, we surveyed the literature to corroborate findings from similar previous studies. However, the risk still remains for oversight or recalibration of model parameters or statistical testing. Moreover, all experiments were conducted in a specific setting; if the setting suffers changes (e.g., involve other domain experts), our results might differ in, for example, number of clusters.

External validity threats relate to the generalizability of the results. The knee OA phenotypes were derived from the CHECK dataset, thus the results are generalizable to a similar population with regards to its characteristics. In terms of the contribution of the MBCFD method, we found that it is superior in the univariate case and very similar in the multivariate case. However, these results depend on the evaluation of domain experts as well as statistical tests with a particular set of features. Working with different experts and applying different tests, might yield different results. In addition, further qualitative and quantitative analysis with a different dataset might facilitate other analytical and statistical generalizations.

Another potential threat is the subject population as inclusion/exclusion criteria of the CHECK study required the subjects to be 45 to 65 years old and ability to understand the Dutch language. The final CHECK cohort is composed of 79% females and 98% Caucasian. Hence, these findings might not be generalizable to a population with a non-Caucasian male majority.

Reliability refers to the consistency of measurements and dependence on particular researchers. We present the premise that if the same experts and researchers were to conduct the experiment again, it would yield similar results. In terms of mitigation strategies, we followed instructions from the software packages and specific seeds were set at the beginning of experiments. Additionally, the same set of subjects/knees and features were used for both clustering exercises. Regarding the semi-systematic literature review, if the PRISMA flowchart were to be replicated, it should produce similar results.

Chapter 5

Conclusion

Osteoarthritis (OA) is the most common form of arthritis; it is a heterogeneous disease characterized by multi-tissue failure in joints and the knee is among the most affected joints. Although the exact cause of OA remains poorly understood, a number of likely relevant and distinct pathobiological and pain features have been identified. The relevance of these mechanisms might vary between patients because distinct phenotypes share distinct underlying mechanisms with different structural and functional consequences. Accordingly, the concept of OA heterogeneity has been gaining renewed interest recently in the pursuit of disease-modifying treatment options. Indeed, there are no effective disease-modifying drugs for knee OA, in large part because clinical trials have treated all knee OA as the same disease, disregarding etiology or risk factors. In this thesis, we used the CRISP-IDM method to investigate whether an MBCFD method can contribute to derive clinically-relevant and statistically-significant knee OA phenotypes and whether this method outperforms traditionally-used HCA. By gaining a better understanding of OA heterogeneity (i.e., finding different phenotypes), we could potentially contribute to the design of clinical trials, prevention strategies, and treatments.

By following the phases and activities of the CRISP-ISM method, we were able to create a set of deliverables that helped us find answers to our research questions. First, we aimed to determine which are the most commonly-used methods in the scientific literature being used to derive knee OA phenotypes and discovered that the most widely used methods are HCA, LCA, k-means, and logistic regression. Therefore, we selected HCA to compare with the MBCFD method. Additionally, we investigated the characteristics or features used in knee OA phenotype studies as well as potential MBCFD algorithms which can be used for the same purpose. We discovered the characteristics can be grouped as pain, radiographic measures (X-ray), clinical measures, biochemical markers, and gene expression. Regarding the potential MBCFD algorithms which can be used to cluster phenotypes, we compared two adaptive model-based clustering methods and decided to choose funHDDC based on its flexibility, interpretability, and ability to handle multivariate data. Through MFPCA and a functional latent mixture model, funHDDC considers the possibility that data can exist in subspaces with different dimensions and the dependency of data points through time. The use of basis function systems allows for the flexible representation of the data as curves and the management of missing data.

Overall, we present two main contributions to the knee OA phenotype field. The first contribution is the finding that an MBCFD algorithm (i.e., funHDDC) was able to detect clinically-relevant and statistically-significant knee OA phenotypes for the univariate case. However, for the multivariate case, the phenotypes were clinically relevant but no statistical significance was found between the groups. The second contribution we found, when comparing an MBCFD method to the widely-used HCA, was that funHDDC outperforms HCA in the univariate case but not in the multivariate case. However, when pursuing clinically-relevant phenotypes during the data exploration phase, we evidenced that achieving a good solution can be complicated and subjective. The main reason behind this challenge is the lack of ground truth labels in unsupervised learning, which is why clustering is typically used in the exploratory stages of a data mining project. Therefore, the contribution of the derived phenotypes should be further investigated in the context of the domain.

5.1 Limitations

The CHECK cohort contains 573 variables. However, due to the nature of functional data analysis, we were limited to data available longitudinally. For example, it would have been interesting to include data on pain per knee but joint-specific pain scales are not available for the 10-year period in the CHECK cohort. Moreover, the CHECK cohort contains some limitations regarding race (98% of subjects were Caucasian) and gender (79% of subjects were female). Thus, we can only generalize results to similar populations.

Regarding the iterative unstructured interviews, we received extensive feedback from domain experts for the funHDDC clusters in comparison to the HCA clusters. Clinical relevance was evaluated with knowledge of previous literature, CHECK, and clinical expertise. However, there is a degree of subjectivity that might cause differences in findings across studies using similar approaches. In addition, most of the feedback was related to decisions about the selection/combination of features, the number of clusters and progression detected, which was also used for the HCA clusters. However, more combinations of features could have been tested as well as parameters for funHDDC and HCA such as other linkage methods and distances.

5.2 Future Work

After completion of this study, we identified interesting opportunities for future research.

- Study how HCA performs with other linkage methods such as single and centroid, as well as other distance such as Gower and Minkowski
- Compare funHDDC performance with latent class analysis and/or latent profile analysis
- Try an ensemble models approach with a mixed data type clustering algorithm and include non-longitudinal features
- Validate clusters in labeled data
- Try consensus clustering with bootstrapping to validate funHDDC clusters
- Investigate the proportion of variance explained for each of the clusters found to understand their contribution to the trajectories and specific differences between the groups.

Appendix A

OA Phenotypes in the Literature

TABLE A.1: Clinical phenotypes, adapted from Deveza, et al. (2017)

Characteristic	Source	Features	Method	Phenotypes
Pain	Cardoso et al., 2016	Pressure pain, heat pain, temporal summation of heat pain, cold pain, and temporal summation of mechanical pain	HCA	1) Low sensitivity to pain 2) Average pain sensitivity 3) High TS of pressure pain 4) Cold pain sensitivity 5) Heat pain sensitivity and TS
	Egsgaard et al., 2015	WOMAC subscales, Lequesne index, QoL (EQ-5D); pain catastrophizing, QST measures (PPT, TS and CPM), CIM, CIIIM, CRP and CRPM	HCA	1) Low sensitivity to pain 2) Early phase sensitization 3) Presence of pain sensitization 4) Presence of pain sensitization and catastrophizing
	Osgood et al., 2015	Pain pressure threshold (knee and elbow), DNIC, low-threshold mechano-receptive function, cold allodynia (knee and elbow)	HCA	1) Peripheral and central sensitization with dysfunctional DNIC 2) None or central sensitization and intact DNIC 3) Peripheral sensitization and dysfunctional DNIC 4) None or peripheral sensitization and intact DNIC
	Frey-Law et al., 2016	Heat thresholds and tolerance, punctate pain intensity, pressure pain thresholds, and noxious heat temporal summation	HCA	1) Low pain sensitivity 2) Average pain sensitivity 3) High temporal summation 4) High heat and pressure pain 5) High punctate pain
Psychological profiles	Cruz-Almeida et al., 2013	Depression, coping strategies (positive and negative), hypervigilance/general reactivity/arousability, dispositional optimism, affect (positive and negative), attention to pain and general anger	HCA	1) High optimism, low negative affect 2) Low positive affect 3) Low optimism 4) Somatic sensitivity/pain hypervigilance
Comorbid symptoms profile	Jenkins and McCoy, 2015	Pain, fatigue and depressive symptoms	HCA	1) High depressive symptoms, low pain 2) Average scores of pain, fatigue and depression
	Murphy et al., 2011	Knee pain, depression, fatigue, sleep problems, and total burden of somatic symptoms	HCA	1) High levels of depression, pain, fatigue, illness burden and sleep problems 2) Intermediate levels of depression, moderate fatigue and illness burden, low pain and sleep problems 3) High levels of sleep problems and low severity of other symptoms
	Hoogeboom et al., 2012	Joint-pain comorbidity (presence vs absence on more than half of the days in the past month)	Pre-defined	1) No joint-pain comorbidities 2) Presence of joint pain comorbidities

Table A.1 continued from previous page

Characteristic	Source	Features	Method	Phenotypes
Clinical characteristics	Knoop et al., 2011	KLG, BMI, muscle strength (mean score of right and left quadriceps and hamstring strength), and depression	k-means	1) Minimal joint disease 2) Strong muscle 3) Non-obese and week muscle 4) Obese and week 5) Depressive
	Kittelson et al., 2016	KLG, BMI, quadriceps strength, palpation tenderness (medial joint line, lateral joint line and pes anserine bursa), pain with patellar grind test, comorbidity status (0-12), number of pain sites (0-13), presence of depression, and pain catastrophizing	LCA	1) Higher levels of comorbidity 2) Higher knee joint sensitivity 3) Higher levels of psychological distress and number of pain sites 4) Mild OA
Knee joint alignment	Iijima et al., 2015	Static varus and varus thrust (4 groups based on presence/absence of each feature)	Pre-defined	1) No varus 2) Varus thrust only 3) Static varus only 4) Static varus and varus thrust
Metabolic profile	Lee et al., 2015	Presence of obesity (BMI >= 27.5 kg/m2) and metabolic abnormality (>= 2 metabolic risk factors)	Pre-defined	1) Metabolically healthy normal weight 2) Metabolically abnormal, normal weight 3) Metabolically healthy obesity 4) Metabolically abnormal obesity
	Sowers et al., 2009	Pain pressure threshold (knee and elbow), DNIC, low-threshold mechano-receptive function, cold allodynia (knee and elbow)	Pre-defined	1) Non-obese, no metabolic clustering 2) Non-obese with metabolic clustering 3) Obese without metabolic clustering 4) Obese with metabolic clustering
Gait parameters	Elbaz et al., 2014	WOMAC subscales, Lequesne index, QoL (EQ-5D); pain catastrophizing, QST measures (PPT, TS and CPM), CIM, CIIIM, CRP and CRPM	k-means cluster analysis followed by CART	1) Stride length <115 2) - Stride length 105-115, or - Stride length 95-105 and cadence >65 3) - Stride length 85-95, or - Stride length 95-105 and cadence <=65, or - Stride length 80-85 and cadence >65 4) - Stride length <=80, or - Stride length 80-85 and cadence <=65
Mechanistic factors	Roze et al., 2016	Metabolic syndrome profile or active and lean profile	Pre-defined	1) Metabolic syndrome 2) Active and lean

TABLE A.2: Imaging phenotypes, adapted from Deveza, et al. (2017)

Characteristic	**Source**	**Features**	**Method**	**Phenotypes**
Knee chondrocalcinosis	Abhishek et al., 2016	Chondrocalcinosis (presence/absence in the index joint)	Pre-defined	1) Absence of chondrocalcinosis 2) Presence of chondrocalcinosis
MRI-detected denuded bone areas (dAB)	Cotofana et al., 2013	dAB presence, location (peripheral vs central), size (<=or >10% of respective cartilage plate) and type (cartilage loss vs intra-chondralosteophyte)	Pre-defined	1) No dAB 2) Peripheral dAB 3) Central dAB
Imaging features (MRI cartilage measures andradiographic features) and clinical symptoms	Waarsing et al., 2015	Quantitative MRI measures of cartilage thickness/volume and dAB; semi-quantitative radiographic scores: KLG,osteophytes, JSN, cysts, sclerosis, chondrocalcinosis and attrition (per compartment for the tibia and femur); WOMAC pain and function,VAS pain (past month and past week), knee baseline symptom status	LCA	1) No dAB independent of KLG 2) Minor dAB (medial compartment) 3) Larger dAB with increasing KLG (lateral compartment) 4) Larger dAB with increasing KLG (medial compartment)
Knee joint compartment	Peat et al., 2012	Presence/absence of radiographic PF joint OA (skyline KLG >= 2 or lateral osteophytes >= 1) and/or TF joint OA (postero-anterior KLG >= 2 or posterior osteophytes >= 1)	Pre-defined	1) Isolated PF OA 2) Combined PF/TF OA 3) Isolated TF OA
	Sharif et al., 2006	Presence/absence of radiographic PF joint OA (one or both PF joints KLG >2 but both medial TF joints KLG <3) and/or TF joint OA (one or both medial TF joints KLG >2 but both PF joints scores <3)	Pre-defined	1) Predominant PF OA 2) Predominant TF OA

TABLE A.3: Laboratory phenotypes, adapted from Deveza et al. (2017)

Characteristic	**Source**	**Features**	**Method**	**Phenotypes**
Biochemical marker patterns	Meulenbelt et al., 2007	Bone markers: s-OC, u-CTX-I; collagen markers: u-CTXII, s-PIIANP, s-COMP, u-TIINE; synovium and inflammation markers: u-Glc-Gal-PYD, s-hsCRP; BMI and age	PCA	1) High CTX-I, CTX-II, osteocalcin, GlceGal-PYD and TIINE 2) High hsCRP and BMI 3) High PIIANP, COMP and age
Inflammatory profile	Siebuhr et al., 2014	Levels of serum hsCRP and serum CRPM	Pre-defined	1) Low hsCRP and CRPM 2) Low CRP and high CRPM 3) High hsCRP and low CRPM 4) High hsCRP and CRPM
Cytokine/ chemokine profile (synovial fluid)	Heard et al., 2013	30 cytokines/chemokines (synovial fluid)	PCA and k-means	1) Group 1 2) Group 2 3) Group 3 (groups not described)
Serum biochemical markers of bone metabolism	Berry et al., 2010a	PINP, osteocalcin, CTX-I, NTX-I and ICTP	Pre-defined	1) Low PINP 2) High PINP 3) Low osteocalcin 4) High osteocalcin
Serum biochemical markers of cartilage metabolism	Berry et al., 2010b	COMP, PIIANP and C2C	Pre-defined	1) Low cartilage biomarkers (COMP, PIIANP and C2C) 2) High cartilage biomarkers
Profile of gene expression in peripheral blood leukocytes	Attur et al., 2011	Cohort	HCA	1) Cytokine overexpressors 2) Cytokine underexpressors

TABLE A.4: OA phenotype research from 2018 to 2020

Characteristic	First author, year	Features	Method	Phenotypes
Pain	Vongsirinavarat et al., 2020	Activity limitation variables: maintaining a standing position, stair climbing time and walking time	Two-step cluster analysis, regression analysis	1) No disability 2) mild disability 3) moderate disability 4) severe disability
Mixed (demographic, clinical, radiographic and biomarkers)	Nelson et al., 2019	Varied (n=73): demographic, clinical, radiographic, biochemical markers	DWD, DiProPerm, k-means, t-SNE, PCA	Progressors and non-progressors
Pain	Carlesso et al., 2019	Psychological factors (pain catastrophizing, depressive symptoms), sleep, WSP, and QST measures of pain sensitization (PPT, TS)	LCA, logistic regression	1) Low-to-moderate proportion of PP sensitivity and facilitated TS 2) Low/absent proportion of PP sensitivity and facilitated TS 3) High proportion of PP sensitivity and moderate proportion of facilitated TS 4) Low proportion of PP sensitivity and high proportion of facilitated TS.
Pain	Glicksberg et al., 2019	Pain and genetic factors	PCA, Bayesian Gaussian mixture model for clustering analysis, regression analysis, logistic association analysis	1) Stable 2) Worsening 3) Progressively worsening
Synovial fluid	Carlson et al., 2019	Synovial fluid (1362 metabolite features)	HCA, PLS-DA, logistic regression, PCA	Four distinct subgroups of donors in early and late stage disease that may be representative of metabolomic synovial phenotypes
Pain	Runhaar et al., 2018	NRS (pain scale)	LCGA	1) Always high pain 2) Always low pain 3) Decreasing pain 4) Fluctuating high pain
Mixed (demographic, clinical, radiographic and biomarkers)	Pan et al., 2020	Baseline blood pressure, glucose, triglycerides and HDL cholesterol; MetS, knee X-ray, WOMAC pain	Group-based trajectory modelling for pain trajectories, multi-nominal logistic regression for analysis.	1) Minimal pain 2) Mild pain 3) Moderate pain

Table A.4 continued from previous page

Characteristic	First author, year	Features	Method	Phenotypes
Mixed (demographic, clinical, and radiographic)	Pan et al., 2019	Sex, BMI, emotional problems, education level, comorbidities, number of painful sites and knee structural pathology	LCA, linear regression	1) High prevalence of emotional problems and low prevalence of structural damage 2) High prevalence of structural damage and low prevalence of emotional problems 3) Low prevalence of emotional problems and low prevalence of structural damage
Quality of life and health-related characteristics	Törmälehto et al., 2019	Health-related quality of life, patient-related characteristics, incidence of knee replacement (KR) and prevalence of pain medication	Group-based trajectory modeling, multinomial logistic regression, Cox regression and generalized estimating equation models	1) No change 2) Rapidly worsening 3) Slowly worsening 4) Improving quality of life
Depression and pain	Rathbun et al., 2020	20-item Center for Epidemiological Studies Depression Scale	LCA	1) Asymptomatic 2) Catatonic 3) Anhedonic 4) Melancholic
Functional capacity	Bieleman et al., 2019	Functional capacity evaluation	LCA	1) Weak giving away 2) Stable and able 3) Strong with decline
Pain	Schiphof et al., 2019	NRS	LCGA	1) Always high pain 2) Always low pain 3) Decreasing pain 4) Fluctuating high pain 5) More knee pain 6) More hip pain
Biomarkers (blood-based)	Zhao et al., 2018	Gene expression data	Support vector machine	1) Group A: degenerative OA with glycosaminoglycan biosynthesis and apoptosis 2) Group B: related to Graft versus host disease and antigen processing and presentation
Demographics, biomechanical	Young-Shand et al., 2020	Demographics (age, gender, BMI), biomechanical severity (gait speed, gait angle, OA severity)	PCA, HCA	1) High functioning males 2) Older stiff-kneed males 3) Slower stiff-kneed females 4) High functioning females
Demographics, pain, comorbidities	Munugoda et al., 2020	Ambulatory activity, body mass index, knee pain, and comorbidities	LCA	1) Normal/overweight participants with higher AA, lower pain and lower comorbidities 2) Overweight participants with lower AA, mild pain and higher comorbidities 3) Obese participants with lower AA, mild pain and higher comorbidities
Gene expression	Soul et al., 2018	2692 differentially expressed genes	HCA	1) Group A 2) Group B

Table A.4 continued from previous page

Characteristic	First author, year	Features	Method	Phenotypes
Demographic, clinical, radiographic	Deveza et al., 2019	WOMAC pain, radiographic JSN, pain duration, TKR family history, obesity	LCGA	1) Stable 2) Moderate cartilage loss 3) Substantial cartilage loss
Macrophage phenotype and gene expression in synovial tissue	Wood et al., 2019	Gene expression of synovial macrophages	tSNE	1) cOA macrophages 2) iOA macrophages
Pain and function	Lee et al., 2018	WOMAC pain, WOMAC function	Group-based trajectory modeling (PROC TRAJ)	1) Lower-Early Improvement 2) Moderate-Early Improvement 3) Higher-Delayed Improvement 4) Higher-No Improvement
Biochemical markers	Karsdal et al., 2019	Biochemical markers of bone, cartilage, soft tissue, synovial metabolism	CART, Poisson regression	1) OA control: high cartilage degradation and synovial inflammation 2) OA control: low cartilage degradation 3) RPOA type-2: high cartilage degradation and synovial inflammation

Appendix B

Data Exploration Steps

TABLE B.1: Iterations of data exploration describing steps taken, outcomes, and participants.

Iteration	Steps taken	Outcome
1	Discussion of dataset: meaning of values, potential inclusion and combinations of features	List of preliminary features
2	Discussion of dataset's characteristics, potential hypotheses, work plan and visualization examples	Recommendation of literature and algorithms to review
3	Cleaning of dataset, discussion of inclusion of features	Dataset ready to be used
4	Preliminary exploration of funHDDC algorithm with four variables and b-splines basis system	Understanding of mechanisms of funHDDC
5	Extension of funHDDC algorithm to use 21 variables (n=824)	With 21 variables, funHDDC suggests using K=3 No meaningful trajectories were found
6	Exploration of different number of basis functions and breaks for b-splines basis system	Understanding of mechanisms of b-splines basis system and implications for our model
7	Exploration of alternatives of data imputation	Choice of linear imputation and keeping observations with at least two data points in time
8	Discussion of descriptive statistics table with 36 baseline characteristics for the 3 clusters of 21 variables	No meaningful insights were found
9	Discussion of Clusters with WOMAC (3) and ICOAP (2) subscales For NRS (only) clusters: ran the K=3:8 solution at knee-level	No meaningful trajectories found
10	Exploration of the influence of data imputation: n=944 vs n=594 subjects for 8 KIDA variables	Data imputation flattens the curves and draws some of the clusters together
11	Discussion of clusters from clinical data with removal of data from patients when they had a knee replacement (n=795), K=3:8	Model suggested using K=6 by BIC score No meaningful trajectories found
12	Discussion of clusters with radiographic data at knee level (n=2,004). K=3:8 explored	Some potentially meaningful trajectories were found, suggesting it could be interesting to perform analysis at knee level, needs further discussion
13	Exploration of pain clusters with NRS variables for left and right knee. K=3:8 explored	Model suggested K=5 by BIC score
14	Presentation of WOMAC pain clusters at patient level (no data available specific to joints)	Model suggested K=8 Similar trajectories found for K=6,7 solutions

Table B.1 continued from previous page

Iteration	Steps taken	Outcome
15	Discussion of WOMAC pain and radiographic (KIDA and scoring) clusters at knee level	Model suggested K=8
16	MFPCA was performed on the original set of 21 variables to evaluate which method to use for feature selection	Expert opinion was regarded as the best choice
17	A different set of radiographic variables (14) were evaluated at subject level (n=893)	Model suggested K=6
18	Exploration of 36 baseline characteristics of subjects in WOMAC pain clusters. In addition, the overlap between WOMAC pain and radiographic clusters was discussed	Radiographic clusters 1,2, and 5 have the largest intersection with WOMAC pain cluters (these are also the largest clusters). No meaningful relation was found
19	Discussion of results of analysis of variance of WOMAC pain clusters	Significant results were found for biomarkers: Leptinengml and BSE with one-way ANOVA
20	The first HCA exercise was performed with average linkage method with a new dataset of posterior probabilities of radiographic and WOMAC pain clusters	No meaningful groups were found, needs further analysis
21	Discussion of clustering exercise for seven radiographic variables at subject level with absolute difference between right and left knee as values	No meaningful trajectories were found
22	Discussion of WOMAC pain and function clusters with data from all ten years including baseline	K=5 solutions was deemed as potentially clinically relevant with five distinct trajectories
23	Evaluation of using different basis systems	Selection of polygonal basis as preferred
24	Discussion of radiographic clusters at knee level with different combinations of KIDA and OA scoring variables: 1) KIDA (no BD) + Scoring OA 2) KIDA (with BD) + Scoring OA 3) KIDA (no BD) 4) KIDA (with BD) 5) KIDA Osteophytes 6) KIDA Joint Space Width 7) KIDA Bone Density 8) Scoring OA 9) KIDA Lateral 10) KIDA Medial	Exercise #1 potentially presents a cluster with progressive MeanLatJSW, further analysis needed Request of calculating the mean between femur and tibia osteophytes to use as new variable Request of calculating the mean between femur and tibia bone density to use as new variable Exercise #8, the K=6 solution is preferred
25	Discussion of WOMAC pain and WOMAC function clusters, as well as their overlap	Meaningful trajectories were found. Potentially interesting overlap results which need further analysis
26	Discussion of MFPCA analysis performed on Exercise #1 from iteration 23	The first principal component explains 43% of the proportion of variance No clear contribution from the variables seen for cluster seven of MeanLatJSW

Table B.1 continued from previous page

Iteration	Steps taken	Outcome
27	Discussion of radiographic clusters at knee level with different combinations of KIDA and OA scoring variables: 1) KIDA + Scoring with BD lateral and medial 2) KIDA + Scoring without BD 3) KIDA with BD lateral and medial 4) KIDA without BD 5) Scoring OA K=6 preferred 6) WOMAC Stiffness	Exercise #3, K=8 is preferred Exercise #6, K=4 then 5 preferred Requested descriptive statistics tables for Exercise #1 from iteration 24
28	Presentation of new analysis with WOMAC stiffness as a new exercise. Computed newly aggregated variables for KIDA analysis with the mean of femur and tibia for osteophytes and bone density.	Selected the following exercises: 1) WOMAC pain K=5 2) WOMAC function K=5 (included overlap chart) 3) WOMAC stiffness K=5 4) KIDA with 4 new variables: osteophytes and BD (lateral and medial), and MeanJSW (lateral and medial) 5) OA Scoring: sclerosis, JSN, and osteophytes Exercise #1 of iteration 24 will no longer be pursued
As a final step, we prepared the HCA clusters and compared to funHDDC results during two sessions.		

Appendix C

Baseline Characteristics for MBCFD and HCA Clusters

TABLE C.1: Baseline characteristics of WOMAC Pain MBCFD clusters

#	Variable	Cluster 1 (n=66)	Cluster 2 (n=77)	Cluster 3 (n=315)	Cluster 4 (n=312)	Cluster 5 (n=48)	p-value
1	Lft_T0	55.05 ± 4.97	55.93 ± 5.57	55.86 ± 5.16	56.22 ± 5.29	55.34 ± 5.3	0.05
2	RAS	98.41	96.05	97.47	98.74	97.87	NA
3	SEXE	79.37	78.95	80.06	78.23	72.34	NA
4	BMI	26.86 ± 4.05	25.58 ± 3.38	26.16 ± 4.09	26.18 ± 4.08	25.06 ± 3.18	0.70
5	Menopauze_01	31 (78)	37 (79)	148 (77)	154 (79)	23 (85)	NA
6	Leptinengml	18.57 ± 18.04	16.03 ± 14.06	17.84 ± 17.67	16.37 ± 15.08	13.32 ± 10.76	0.90
7	Adiponectineugml	11.08 ± 5.56	11.98 ± 6.12	12 ± 6.76	12.02 ± 7.34	11.35 ± 5.2	0.44
8	Resistinengml	3.76 ± 1.38	3.91 ± 1.41	3.82 ± 1.42	3.75 ± 1.08	3.28 ± 0.89	0.82
9	CTXIugmmol	186.23 ± 132.63	180.56 ± 134.78	180.32 ± 115.76	178.77 ± 113.54	184.39 ± 120.6	0.56
10	uNTXInMBCEmmol	43.11 ± 19.81	41.45 ± 19.85	41.76 ± 20.23	41.71 ± 21.1	39.75 ± 15.81	0.28
11	sPINP	43.01 ± 15.65	46.8 ± 20.06	46.27 ± 19.33	45.11 ± 19.31	49.22 ± 25.42	0.30
12	sOC	13.67 ± 4.13	14.65 ± 6.83	14.58 ± 7.58	14.08 ± 5.75	14.22 ± 5.75	0.84
13	sC12C	0.28 ± 0.6	0.19 ± 0.07	0.2 ± 0.26	0.23 ± 0.46	0.2 ± 0.12	0.02
14	CTXIIngmmol	206.28 ± 118.27	211.92 ± 122.92	223.58 ± 120.13	229.96 ± 139.86	198.62 ± 106.39	0.48
15	sCS846	76.56 ± 57.48	71.24 ± 33.51	75.7 ± 40.6	82.03 ± 60.41	78.36 ± 44.95	0.47
16	sCOMPUl	8.34 ± 1.38	8.72 ± 2.55	8.77 ± 2.48	8.78 ± 2.16	9.14 ± 2.5	0.43
17	sPIIANP	1408.6 ± 382.87	1414.75 ± 594.83	1488.18 ± 610.74	1597.82 ± 854.89	1505.97 ± 622.34	0.92
18	sHA	30.83 ± 21.4	30.76 ± 18.84	33.11 ± 28.47	35.27 ± 25.44	32.3 ± 31.01	0.91
19	sPIIINP	4.23 ± 1.11	4.24 ± 1.06	4.29 ± 1.39	4.33 ± 1.15	4.45 ± 1.03	0.98
20	hsCRP	2.68 ± 3.94	3.66 ± 6.4	3.09 ± 6.5	2.62 ± 3.29	2.44 ± 3.54	0.70
21	BSE	10.32 ± 7.91	9.74 ± 7.16	10.31 ± 8.18	9.46 ± 6.52	9.39 ± 7.37	0.33

TABLE C.2: Baseline characteristics of WOMAC Function MBCFD clusters

#	Variable	Cluster 1 (n=80)	Cluster 2 (n=274)	Cluster 3 (n=37)	Cluster 4 (n=323)	Cluster 5 (n=105)	p-value
1	Lft_T0	56.21 ± 5.51	56.11 ± 5.19	56.3 ± 5.01	55.72 ± 5.16	55.55 ± 5.72	0.80
2	RAS	95	98.19	100	98.14	97.06	NA
3	SEXE	85	79.78	75.68	77.09	77.45	NA
4	BMI	25.9 ± 4.31	26.15 ± 4.14	26.73 ± 4.29	26.21 ± 3.9	25.76 ± 3.55	0.52
5	Menopauze_01	43 (83)	135 (78)	16 (76)	148 (77)	49 (78)	NA
6	Leptinengml	17.42 ± 16.52	16.43 ± 15.48	17.53 ± 17.55	17.92 ± 17.06	14.36 ± 12.88	0.76
7	Adiponectineugml	12.46 ± 7.09	11.74 ± 6.31	11.82 ± 6.11	11.72 ± 7.18	12.29 ± 6.66	0.54
8	Resistinengml	3.68 ± 1.02	3.87 ± 1.46	3.8 ± 1.19	3.73 ± 1.24	3.63 ± 1	0.43
9	CTXIugmmol	180.23 ± 95.47	169.21 ± 98.81	172.16 ± 79.73	187.36 ± 127.64	190.4 ± 156.73	0.58
10	uNTXInMBCEmmol	42.59 ± 18.66	41.24 ± 21.95	41.76 ± 14.65	41.21 ± 19.5	42.8 ± 20.48	0.68
11	sPINP	43.18 ± 14.93	44.87 ± 20.99	44.18 ± 14.25	47.73 ± 19.9	45.25 ± 19.38	0.13
12	sOC	13.42 ± 4.78	14.12 ± 7.49	13.81 ± 4.37	14.62 ± 5.97	14.71 ± 7.02	0.17
13	sC12C	0.32 ± 0.8	0.23 ± 0.45	0.16 ± 0.05	0.19 ± 0.1	0.18 ± 0.07	0.15
14	CTXIIngmmol	240.3 ± 173.38	216.19 ± 119.49	238.21 ± 136.36	226.34 ± 123.01	205.56 ± 116.28	0.50
15	sCS846	87.15 ± 79.66	76.72 ± 44.68	88.38 ± 44.75	77.02 ± 51.39	73.32 ± 34.61	0.30
16	sCOMPUl	8.87 ± 2.17	8.71 ± 2.39	8.83 ± 1.86	8.85 ± 2.37	8.56 ± 2.28	0.81
17	sPIIANP	1536.68 ± 570.26	1528.22 ± 848.74	1721.94 ± 1029.74	1513.2 ± 571.28	1368.4 ± 524.07	0.05
18	sHA	36.49 ± 28.8	33.49 ± 23.74	37.62 ± 25.3	32.65 ± 28.2	33.84 ± 25.39	0.70
19	sPIIINP	4.41 ± 1.32	4.27 ± 1.06	4.33 ± 1.18	4.37 ± 1.39	4.12 ± 1.04	0.71
20	hsCRP	2.35 ± 2.78	2.9 ± 4.64	3.09 ± 4.18	3.25 ± 6.46	2.24 ± 2.68	0.41
21	BSE	9.05 ± 6.17	10.13 ± 8.19	11.29 ± 7.37	9.9 ± 7.21	9.53 ± 6.5	0.54

TABLE C.3: Baseline characteristics of WOMAC Stiffness MBCFD clusters

#	Variable	Cluster 1 (n=325)	Cluster 2 (n=91)	Cluster 3 (n=89)	Cluster 4 (n=126)	Cluster 5 (n=189)	p-value
1	Lft_T0	56.16 ± 5.22	55.45 ± 5.19	56.18 ± 4.97	55.08 ± 5.36	56.11 ± 5.42	0.30
2	RAS	97.53	97.75	98.88	96.83	98.44	NA
3	SEXE	81.17	77.53	79.78	76.19	76.56	NA

Table C.3 continued from previous page

#	Variable	Cluster 1 (n=325)	Cluster 2 (n=91)	Cluster 3 (n=89)	Cluster 4 (n=126)	Cluster 5 (n=189)	p-value
4	BMI	26.15 ± 4.13	26.22 ± 3.87	25.85 ± 4.01	25.98 ± 3.69	26.28 ± 4.05	0.90
5	Menopauze_01	156 (76)	44 (77)	40 (78)	57 (77)	95 (81)	NA
6	Leptinengml	17.45 ± 17.07	16.6 ± 15.13	15.5 ± 13.75	17.15 ± 15.34	16.63 ± 16.17	0.97
7	Adiponectineugml	11.91 ± 6.79	11.82 ± 6.58	11.72 ± 6.65	11.54 ± 6.42	12.12 ± 7.13	0.99
8	Resistinengml	3.65 ± 1.06	3.57 ± 0.99	3.99 ± 1.5	3.74 ± 1.11	3.94 ± 1.63	0.17
9	CTXIugmmol	183.81 ± 136.59	159.7 ± 90.89	194.23 ± 122.53	178.24 ± 91.54	179.11 ± 110.48	0.33
10	uNTXInMBCEmmol	41.32 ± 22.37	37.46 ± 15.35	44.12 ± 21.42	41.55 ± 16.27	42.84 ± 20.03	0.14
11	sPINP	44.34 ± 17.73	40.98 ± 15.14	48.06 ± 21.01	47.02 ± 20.29	48.91 ± 22.61	0.05
12	sOC	14.15 ± 6.77	13.24 ± 4.71	14.58 ± 6.15	14.43 ± 6.2	14.89 ± 7.1	0.38
13	sC12C	0.19 ± 0.08	0.19 ± 0.1	0.27 ± 0.61	0.2 ± 0.12	0.26 ± 0.62	0.27
14	CTXIIngmmol	219.49 ± 120.56	215.33 ± 118.03	242.09 ± 158.26	226.03 ± 141.11	218.4 ± 118.11	0.81
15	sCS846	76.71 ± 42.31	83.47 ± 55.55	78.81 ± 71.87	79.18 ± 63.98	76.15 ± 38.94	0.50
16	sCOMPUl	8.54 ± 2.23	9.12 ± 2.51	8.85 ± 2.77	8.6 ± 2.04	9.07 ± 2.33	0.09
17	sPIIANP	1537.12 ± 872.35	1581.17 ± 619.01	1473.33 ± 598.5	1510.3 ± 500.84	1453.29 ± 542.23	0.48
18	sHA	33.67 ± 24.69	33.43 ± 25.96	33.31 ± 21.14	36.31 ± 35.88	32.18 ± 23.67	0.93
19	sPIIINP	4.29 ± 1.17	4.27 ± 1.01	4.36 ± 1.21	4.42 ± 1.54	4.27 ± 1.19	0.90
20	hsCRP	3.26 ± 6.6	2.88 ± 3.89	2 ± 2.67	2.57 ± 3.93	2.96 ± 4.23	0.47
21	BSE	11.01 ± 8.26	9.67 ± 7.69	7.95 ± 5.71	9.3 ± 6.88	9.47 ± 6.37	0.01

TABLE C.4: Baseline characteristics of KIDA MBCFD clusters

#	Variable	Cluster 1 (n=378)	Cluster 2 (n=283)	Cluster 3 (n=199)	Cluster 4 (n=134)	Cluster 5 (n=61)	Cluster 6 (n=407)	Cluster 7 (n=99)	Cluster 8 (n=227)	p-value
1	Lft_T0	55.93 ± 5.38	55.76 ± 5.11	55.91 ± 5.02	57.15 ± 5.07	56.85 ± 5.94	55.78 ± 5.22	55.17 ± 4.76	55.92 ± 5.34	0.46
2	RAS	98.68	98.23	94.97	97.76	96.72	98.03	98.99	97.36	NA
3	SEXE	75.13	82.33	80.4	78.36	75.41	81.57	75.76	77.09	NA
4	BMI	26.15 ± 4.22	25.84 ± 3.93	25.57 ± 3.85	26.29 ± 3.84	26.25 ± 3.97	26.35 ± 3.85	25.92 ± 4.24	26.29 ± 3.95	0.81
5	Menopauze_01	174 (74)	136 (77)	96 (77)	72 (89)	31 (82)	189 (77)	40 (73)	106 (77)	NA
6	Leptinengml	17.21 ± 17.05	16.82 ± 14.67	15.34 ± 13.58	16.02 ± 14.77	14.73 ± 13.8	18.56 ± 17.62	14.27 ± 14.45	17.59 ± 17.14	0.93
7	Adiponectineugml	12.11 ± 7.29	11.82 ± 6.35	12.14 ± 6.59	13.19 ± 6.73	10.21 ± 6.81	11.57 ± 6.42	12.52 ± 7.48	11.03 ± 6.01	0.18
8	Resistinengml	3.81 ± 1.15	3.63 ± 1.09	3.84 ± 1.35	3.66 ± 1.23	3.56 ± 1.08	3.8 ± 1.16	3.79 ± 1.27	3.88 ± 1.79	0.08
9	CTXIugmmol	175.41 ± 121.34	184.49 ± 114.72	180.39 ± 115.64	182.67 ± 117.73	161.72 ± 123.62	178.37 ± 108.22	188.11 ± 127.03	178.25 ± 134.78	0.63
10	uNTXInMBCEmmol	40.57 ± 18.83	43.8 ± 22.1	42.14 ± 22.55	45.52 ± 18.32	38.34 ± 20.23	40.29 ± 19.75	41.56 ± 19.42	38.75 ± 19.07	0.46

Table C.4 continued from previous page

#	Variable	Cluster 1 (n=378)	Cluster 2 (n=283)	Cluster 3 (n=199)	Cluster 4 (n=134)	Cluster 5 (n=61)	Cluster 6 (n=407)	Cluster 7 (n=99)	Cluster 8 (n=227)	p-value
11	sPINP	46.36 ± 20.47	46.05 ± 17.95	45.29 ± 19.69	46.7 ± 18.87	39.86 ± 14.73	46.26 ± 21.62	49.44 ± 25.92	44.31 ± 21.35	0.96
12	sOC	14.08 ± 6.67	14.33 ± 6.7	14.9 ± 6.97	14.83 ± 5.43	12.99 ± 4.52	14.28 ± 5.93	14.71 ± 8.43	14.04 ± 7.41	0.75
13	sC12C	0.19 ± 0.1	0.24 ± 0.51	0.26 ± 0.65	0.23 ± 0.49	0.17 ± 0.06	0.21 ± 0.3	0.24 ± 0.57	0.21 ± 0.33	0.39
14	CTXIIngmmol	218.36 ± 127.64	224.98 ± 131.47	217.78 ± 124.43	241.05 ± 132.41	219.6 ± 119.47	225.15 ± 133.86	212.56 ± 109.42	213.96 ± 122.68	0.47
15	sCS846	79.19 ± 52.34	75.3 ± 31.67	79.36 ± 50.85	78.49 ± 33.6	72.18 ± 44.56	76.5 ± 48.68	75.73 ± 36.1	83.25 ± 73.79	0.96
16	sCOMPUl	8.86 ± 2.39	8.47 ± 2.16	8.61 ± 2.4	8.9 ± 2.25	8.62 ± 2.27	8.81 ± 2.34	8.41 ± 2.32	8.79 ± 2.1	0.91
17	sPIIANP	1499.07 ± 722.25	1447.04 ± 591.1	1535.6 ± 808.19	1513.24 ± 608.02	1297.33 ± 460.71	1520.72 ± 680.21	1552.32 ± 573.37	1609.47 ± 904.72	0.27
18	sHA	33.02 ± 25.2	33.33 ± 24.08	31.79 ± 21.48	35.51 ± 20.03	28.73 ± 18.32	32.89 ± 23	36.87 ± 46.88	33.93 ± 29.57	0.37
19	sPIIINP	4.35 ± 1.3	4.28 ± 1.1	4.28 ± 1.12	4.31 ± 1.04	4.23 ± 0.88	4.34 ± 1.15	4.47 ± 2.14	4.16 ± 1.04	0.78
20	hsCRP	2.72 ± 3.94	2.69 ± 3.67	3.25 ± 10.11	2.07 ± 2.19	2.93 ± 2.96	3.21 ± 5.11	3.18 ± 6.39	2.93 ± 4.68	0.56
21	BSE	9.67 ± 8.2	10.42 ± 6.89	8.96 ± 6.99	8.87 ± 6.92	8.95 ± 5.99	10.53 ± 7.64	9.95 ± 6.86	9.27 ± 6.82	0.66

TABLE C.5: Baseline characteristics of OA Scoring MBCFD clusters

#	Variable	Cluster 1 (n=517)	Cluster 2 (n=259)	Cluster 3 (n=292)	Cluster 4 (n=409)	Cluster 5 (n=81)	Cluster 6 (n=230)	p-value
1	Lft_T0	55.97 ± 5.1	56.35 ± 5.12	55.82 ± 5.35	55.56 ± 5.17	55.9 ± 5.5	56.3 ± 5.48	0.53
2	RAS	98.84	96.14	96.92	98.53	96.3	97.39	NA
3	SEXE	77.76	76.06	81.85	81.42	70.37	79.13	NA
4	BMI	26.12 ± 3.9	26.39 ± 3.77	25.67 ± 3.62	25.98 ± 4.08	26.42 ± 4.15	26.39 ± 4.53	0.72
5	Menopauze_01	252 (78)	118 (80)	143 (76)	190 (73)	40 (87)	101 (81)	NA
6	Leptinengml	17.34 ± 17.05	18.07 ± 16.46	16.07 ± 15.01	15.99 ± 14.58	17.21 ± 16.8	17.65 ± 17.09	0.27
7	Adiponectineugml	11.7 ± 6.77	11.43 ± 6.12	12.11 ± 6.51	12 ± 6.81	10.84 ± 5.6	12.38 ± 7.39	0.79
8	Resistinengml	3.72 ± 1.3	3.81 ± 1.26	3.69 ± 1.16	3.85 ± 1.17	3.66 ± 1.21	3.84 ± 1.55	0.79
9	CTXIugmmol	175.24 ± 116.03	185.5 ± 121.92	173.51 ± 102.6	183.89 ± 128.58	186.45 ± 122.86	177.55 ± 120.54	0.86
10	uNTXInMBCEmmol	40.83 ± 20.03	42.26 ± 20.72	41.35 ± 20.03	42.08 ± 21.05	40.63 ± 15.98	40.22 ± 19.78	0.92
11	sPINP	44.67 ± 20.92	46.37 ± 19.86	47.45 ± 18.71	46.2 ± 20.52	44.46 ± 15.81	46 ± 23.43	0.71
12	sOC	14.06 ± 6.65	14.53 ± 6.82	14.32 ± 5.44	14.23 ± 6.7	14.54 ± 6.95	14.61 ± 7.32	0.73
13	sC12C	0.21 ± 0.37	0.21 ± 0.36	0.24 ± 0.5	0.23 ± 0.43	0.19 ± 0.07	0.22 ± 0.39	0.10
14	CTXIIngmmol	221.25 ± 127.87	236.03 ± 141.56	213.07 ± 118.82	230.57 ± 128.5	221.52 ± 140.18	202.84 ± 114.28	0.71
15	sCS846	79.86 ± 65.09	74.33 ± 29.11	77.68 ± 47.78	76.11 ± 37.19	79.86 ± 30.74	80.86 ± 55.16	0.42
16	sCOMPUl	8.79 ± 2.31	8.62 ± 2.12	8.68 ± 2.21	8.77 ± 2.56	8.64 ± 2.28	8.66 ± 2.08	0.08
17	sPIIANP	1444.64 ± 642.76	1558.31 ± 729.2	1626.28 ± 894.72	1469.98 ± 698.75	1492.32 ± 545.2	1536.4 ± 601.22	0.90
18	sHA	31.59 ± 23.6	38.35 ± 34.95	32.23 ± 21.76	32.32 ± 24.07	39.92 ± 24.26	31.96 ± 26.85	0.60
19	sPIIINP	4.27 ± 1.14	4.35 ± 1.62	4.35 ± 1.1	4.33 ± 1.13	4.16 ± 1.12	4.29 ± 1.19	0.28
20	hsCRP	3.02 ± 6.39	2.9 ± 3.8	2.55 ± 4.57	2.82 ± 5.86	2.47 ± 3.26	3.35 ± 4.43	0.49
21	BSE	9.45 ± 7.65	10.09 ± 7.11	9.44 ± 6.76	9.89 ± 7.42	10.14 ± 8.27	10.28 ± 7.15	0.98

TABLE C.6: Baseline characteristics WOMAC Pain HCA clusters with Ward linkage method

#	Variable	Cluster 1 (n=92)	Cluster 2 (n=248)	Cluster 3 (n=200)	Cluster 4 (n=136)	Cluster 5 (n=143)	p-value
1	Lft_T0	56.07 ± 5.67	55.93 ± 5.21	56.03 ± 5.01	55.8 ± 5.33	55.67 ± 5.44	0.97
2	RAS	95.65	97.58	98	100	97.2	NA
3	SEXE	77.17	77.02	79	82.35	79.02	NA
4	BMI	26.26 ± 4.67	26.03 ± 3.96	26.26 ± 4.16	26.3 ± 3.73	25.87 ± 3.63	0.80
5	Menopauze_01	44 (73)	114 (78)	89 (74)	76 (85)	68 (79)	NA
6	Leptinengml	17.97 ± 17.98	15.65 ± 14.64	18.25 ± 17.92	17.22 ± 15.21	16.23 ± 14.97	0.77
7	Adiponectineugml	11.85 ± 6.89	11.25 ± 6.14	12.34 ± 7.25	12.59 ± 7.44	11.61 ± 6.32	0.63
8	Resistinengml	3.72 ± 1.11	3.79 ± 1.52	3.84 ± 1.19	3.75 ± 1.22	3.63 ± 1.05	0.53
9	CTXIugmmol	186.8 ± 135.1	171.17 ± 98.73	187.02 ± 141.05	174.99 ± 87.7	187.17 ± 128.57	0.89
10	uNTXInMBCEmmol	41.57 ± 18.25	41.62 ± 22.85	41.69 ± 20.07	41.15 ± 16.59	41.76 ± 20.05	0.96
11	sPINP	43.94 ± 17.47	45.12 ± 21.43	46.53 ± 21.28	47.25 ± 17.05	46.09 ± 17.43	0.27
12	sOC	13.79 ± 5.05	14.04 ± 6.49	14.74 ± 7.91	13.98 ± 4.78	14.82 ± 6.62	0.33
13	sC12C	0.24 ± 0.46	0.24 ± 0.47	0.18 ± 0.07	0.19 ± 0.12	0.23 ± 0.48	0.18
14	CTXIIngmmol	217.48 ± 156.38	218.43 ± 126.05	217.96 ± 112.08	223.84 ± 119.84	235.59 ± 136.43	0.41
15	sCS846	76.2 ± 31.55	72.48 ± 30.66	80.6 ± 52.92	80.14 ± 65.63	82.6 ± 66.27	0.46
16	sCOMPUl	8.45 ± 2.3	8.71 ± 2.39	8.83 ± 2.37	9.14 ± 2.4	8.65 ± 2.09	0.23
17	sPIIANP	1500.35 ± 660.32	1523.51 ± 799.45	1542.13 ± 718.66	1550.16 ± 628.13	1419.07 ± 556.6	0.35
18	sHA	39.34 ± 34.84	32.94 ± 22.21	33.64 ± 24.66	32.58 ± 21.58	32.37 ± 32.04	0.53
19	sPIIINP	4.18 ± 1.15	4.26 ± 1.08	4.32 ± 1.09	4.41 ± 1.33	4.37 ± 1.55	0.64
20	hsCRP	2.66 ± 3.98	2.87 ± 4.46	2.96 ± 4.29	2.94 ± 3.67	3.05 ± 8.19	0.61
21	BSE	8.97 ± 5.7	10.64 ± 8.61	9.7 ± 7.18	9.71 ± 7.04	9.75 ± 6.66	0.89

TABLE C.7: Baseline characteristics WOMAC Function HCA clusters with Ward linkage method

#	Variable	Cluster 1 (n=69)	Cluster 2 (n=186)	Cluster 3 (n=188)	Cluster 4 (n=202)	Cluster 5 (n=174)	p-value
1	Lft_T0	55.58 ± 5.75	55.69 ± 5.03	55.94 ± 5.28	56.06 ± 5.34	56.03 ± 5.26	0.92
2	RAS	95.65	97.85	98.4	98.02	97.7	NA
3	SEXE	78.26	77.96	81.91	79.21	75.86	NA

Table C.7 continued from previous page

#	Variable	Cluster 1 (n=69)	Cluster 2 (n=186)	Cluster 3 (n=188)	Cluster 4 (n=202)	Cluster 5 (n=174)	p-value
4	BMI	25.81 ± 4.06	25.89 ± 3.96	26.18 ± 3.93	26.28 ± 4.04	26.3 ± 4.07	0.64
5	Menopauze_01	34 (79)	87 (80)	88 (71)	101 (80)	81 (82)	NA
6	Leptinengml	17.44 ± 19.87	15.7 ± 15.14	17.04 ± 15.04	16.97 ± 15.46	17.77 ± 17.05	0.58
7	Adiponectineugml	12.55 ± 7.59	11.84 ± 6.46	11.47 ± 6.31	11.93 ± 6.96	12 ± 7.03	0.90
8	Resistinengml	3.74 ± 1.3	3.76 ± 1.56	3.91 ± 1.31	3.66 ± 1.06	3.71 ± 1.08	0.47
9	CTXIugmmol	178.44 ± 87.35	172.73 ± 98.81	167.84 ± 101.12	201.93 ± 146.55	176.84 ± 125.68	0.14
10	uNTXInMBCEmmol	42.28 ± 17.87	41.49 ± 23.35	41.16 ± 20.66	43.26 ± 19.15	39.9 ± 18.15	0.24
11	sPINP	43.53 ± 13.6	44.37 ± 20.42	45.76 ± 21.09	47.5 ± 18.88	46.48 ± 19.89	0.22
12	sOC	13.98 ± 4.88	13.69 ± 6.39	14.5 ± 7.7	14.76 ± 6	14.37 ± 6.36	0.21
13	sC12C	0.19 ± 0.09	0.25 ± 0.55	0.19 ± 0.07	0.24 ± 0.51	0.19 ± 0.09	0.64
14	CTXIIngmmol	225.21 ± 160.55	223.98 ± 127.14	215.31 ± 111.19	233.16 ± 134.11	213.58 ± 121.74	0.76
15	sCS846	77.02 ± 30.08	73.21 ± 31.39	81.73 ± 53.16	79.22 ± 57.72	77.66 ± 60.97	0.43
16	sCOMPUl	8.74 ± 2.18	8.74 ± 2.46	8.77 ± 2.25	8.78 ± 2.35	8.78 ± 2.33	0.97
17	sPIIANP	1582.05 ± 710.88	1545.22 ± 960.55	1526.75 ± 595.99	1494.99 ± 623.76	1451.52 ± 536.74	0.74
18	sHA	35.02 ± 28.78	33.4 ± 22.99	32.72 ± 26.07	35.95 ± 27.63	31.79 ± 27.32	0.66
19	sPIIINP	4.45 ± 1.21	4.28 ± 1.14	4.29 ± 1.07	4.26 ± 1.32	4.37 ± 1.37	0.60
20	hsCRP	2.61 ± 3.65	2.83 ± 4.84	3.78 ± 8.01	2.61 ± 3.52	2.54 ± 3.07	0.36
21	BSE	9.48 ± 6.76	9.98 ± 8.51	10.71 ± 7.84	9.35 ± 6.4	9.77 ± 6.93	0.60

TABLE C.8: Baseline characteristics WOMAC Stiffness HCA clusters with Ward linkage method

#	Variable	Cluster 1 (n=113)	Cluster 2 (n=240)	Cluster 3 (n=164)	Cluster 4 (n=256)	Cluster 5 (n=47)	p-value
1	Lft_T0	55.44 ± 5.3	56.2 ± 5.49	56.01 ± 5.29	55.8 ± 5.03	55.72 ± 5.29	0.82
2	RAS	98.23	98.33	96.95	97.27	100	NA
3	SEXE	79.65	77.92	76.22	80.86	78.72	NA
4	BMI	26.29 ± 3.94	26.36 ± 4.11	25.94 ± 3.91	25.99 ± 3.99	25.96 ± 3.98	0.75
5	Menopauze_01	51 (72)	112 (77)	80 (82)	129 (79)	20 (77)	NA
6	Leptinengml	17.1 ± 15.48	16.98 ± 15.77	16.98 ± 15.32	16.76 ± 17.06	16.59 ± 16.04	0.96

Table C.8 continued from previous page

#	Variable	Cluster 1 (n=113)	Cluster 2 (n=240)	Cluster 3 (n=164)	Cluster 4 (n=256)	Cluster 5 (n=47)	p-value
7	Adiponectineugml	11.06 ± 6.38	12.07 ± 7.23	12.17 ± 6.85	11.97 ± 6.57	11.25 ± 6.01	0.59
8	Resistinengml	3.77 ± 1.13	3.87 ± 1.55	3.67 ± 1.15	3.72 ± 1.13	3.73 ± 1.15	0.67
9	CTXIugmmol	172.44 ± 89.33	178.33 ± 121.05	180.51 ± 124.74	187.68 ± 128.8	166.84 ± 76.14	0.97
10	uNTXInMBCEmmol	40.79 ± 16.07	41.75 ± 19.22	40.39 ± 19.83	42.96 ± 23.72	39.06 ± 12.71	0.90
11	sPINP	45.45 ± 17.83	46.55 ± 20.45	45.26 ± 20.73	46.21 ± 19.66	43.19 ± 14.65	0.95
12	sOC	13.59 ± 5.29	14.46 ± 6.53	14.08 ± 5.18	14.67 ± 7.83	14.12 ± 4.93	0.81
13	sC12C	0.26 ± 0.55	0.24 ± 0.55	0.19 ± 0.09	0.19 ± 0.09	0.21 ± 0.12	0.28
14	CTXIIngmmol	230.26 ± 150.55	223.77 ± 129.64	216.19 ± 118.86	221.63 ± 120.17	217.82 ± 127.23	0.98
15	sCS846	78.15 ± 62.36	76.52 ± 39.87	76.4 ± 34.32	76.99 ± 48.79	95.62 ± 103	0.52
16	sCOMPUl	8.62 ± 2.3	8.93 ± 2.31	8.9 ± 2.33	8.65 ± 2.38	8.46 ± 2.21	0.38
17	sPIIANP	1443.1 ± 493.27	1494.77 ± 719.45	1577.51 ± 774.91	1504.32 ± 731.72	1570.55 ± 504.08	0.58
18	sHA	34.13 ± 34.82	33.94 ± 24.82	35.31 ± 24.36	31.37 ± 23.07	38.24 ± 33.13	0.21
19	sPIIINP	4.41 ± 1.6	4.23 ± 1.19	4.28 ± 1.2	4.35 ± 1.1	4.34 ± 1.17	0.81
20	hsCRP	2.59 ± 4.15	2.99 ± 4.21	2.83 ± 3.74	3.2 ± 7.06	1.96 ± 1.91	0.86
21	BSE	8.52 ± 6.48	10.2 ± 7.55	9.22 ± 6.22	10.84 ± 8.35	9.33 ± 6.36	0.11

TABLE C.9: Baseline characteristics of KIDA HCA clusters with Ward linkage method

#	Variable	Cluster 1 (n=261)	Cluster 2 (n=131)	Cluster 3 (n=63)	Cluster 4 (n=277)	Cluster 5 (n=418)	Cluster 6 (n=529)	Cluster 7 (n=91)	Cluster 8 (n=18)	p-value
1	Lft_T0	55.49 ± 5.02	56.39 ± 5.18	55.89 ± 5.82	55.92 ± 4.91	56.07 ± 5.35	55.89 ± 5.3	56.23 ± 5.63	57.11 ± 4.17	0.29
2	RAS	96.55	97.71	96.83	98.19	98.33	98.11	96.7	94.44	NA
3	SEXE	79.31	79.39	77.78	76.17	79.19	79.58	80.22	77.78	NA
4	BMI	25.85 ± 3.94	25.95 ± 4.09	26.79 ± 3.76	26.3 ± 3.8	26.25 ± 3.98	26.07 ± 4.18	25.4 ± 3.54	26.33 ± 3.36	0.23
5	Menopauze_01	119 (76)	59 (80)	34 (87)	123 (78)	196 (74)	252 (76)	50 (83)	11 (100)	NA
6	Leptinengml	16.05 ± 14.82	15.45 ± 15.72	16.95 ± 13.7	18 ± 16.91	17.4 ± 16.32	17.56 ± 17	13.24 ± 11.96	16.83 ± 16.39	0.47
7	Adiponectineugml	11.92 ± 6.75	11.18 ± 5.75	11.14 ± 5.23	11.73 ± 7.46	12.15 ± 6.83	11.83 ± 6.42	12.37 ± 7.3	10.46 ± 4.59	0.76
8	Resistinengml	3.75 ± 1.3	3.65 ± 1.09	3.59 ± 1.16	3.84 ± 1.36	3.81 ± 1.08	3.79 ± 1.4	3.52 ± 1.21	4.06 ± 1.56	0.63
9	CTXIugmmol	188.96 ± 132.88	164.71 ± 94.27	150.54 ± 82.59	191.43 ± 114.28	181.14 ± 119.82	174.45 ± 116.72	174.62 ± 142.98	176.63 ± 109.49	0.97
10	uNTXInMBCEmmol	43.93 ± 22.43	37.89 ± 15.92	39.92 ± 14.6	43.22 ± 22.4	40.84 ± 18.84	40.66 ± 20.22	39.1 ± 19.44	45.81 ± 19.97	0.88
11	sPINP	46.47 ± 21.69	43.94 ± 22.97	40.79 ± 14.12	48.16 ± 21.31	47.26 ± 22.62	44.88 ± 18.1	44.03 ± 16.53	41.41 ± 16.1	0.56
12	sOC	14.45 ± 6.81	13.54 ± 6.29	12.76 ± 6.17	14.69 ± 5.86	14.66 ± 7.68	14.14 ± 6.33	14.23 ± 5.52	14.11 ± 4.18	0.92
13	sC12C	0.25 ± 0.58	0.19 ± 0.1	0.18 ± 0.07	0.22 ± 0.44	0.22 ± 0.3	0.21 ± 0.28	0.3 ± 0.82	0.18 ± 0.07	0.83
14	CTXIIngmmol	220.05 ± 129.35	220.39 ± 120.13	223.02 ± 90.96	222.71 ± 150.04	222.17 ± 117.91	220.51 ± 126.29	221.42 ± 129.31	265.26 ± 164.23	0.85
15	sCS846	79.7 ± 48.09	84.9 ± 59.55	70.14 ± 22	75.19 ± 30.39	77.72 ± 42.52	78.7 ± 63.3	73.21 ± 41.39	84.5 ± 49.99	0.46

Table C.9 continued from previous page

#	Variable	Cluster 1 (n=261)	Cluster 2 (n=131)	Cluster 3 (n=63)	Cluster 4 (n=277)	Cluster 5 (n=418)	Cluster 6 (n=529)	Cluster 7 (n=91)	Cluster 8 (n=18)	p-value
16	sCOMPUl	8.66 ± 2.27	8.66 ± 2.2	8.63 ± 2.32	8.58 ± 2.11	8.76 ± 2.38	8.86 ± 2.33	8.53 ± 2.44	8.42 ± 1.94	0.33
17	sPIIANP	1485.63 ± 828.13	1632.48 ± 804.61	1455.52 ± 553.33	1479.01 ± 571.62	1538.69 ± 628.7	1525.13 ± 788.04	1348.6 ± 510.49	1432.74 ± 491.27	0.33
18	sHA	32.24 ± 21.14	34.56 ± 19.93	30.71 ± 17.56	31.88 ± 24.45	33.04 ± 29.94	35.42 ± 28.77	29.52 ± 17.34	31.99 ± 23.56	0.27
19	sPIIINP	4.24 ± 1.15	4.41 ± 1.11	4.43 ± 1.12	4.33 ± 1.15	4.35 ± 1.44	4.28 ± 1.14	4.18 ± 1.21	4.08 ± 0.69	0.49
20	hsCRP	3.06 ± 8.88	2.44 ± 3.08	2.88 ± 3.43	2.91 ± 4.82	2.96 ± 4.99	2.87 ± 4.29	2.84 ± 3.99	3.26 ± 3.73	0.12
21	BSE	9.48 ± 7.11	8.5 ± 6.34	9.63 ± 7.46	9.17 ± 7.04	10.64 ± 7.48	9.72 ± 7.66	9.92 ± 6.63	13.82 ± 9.84	0.66

TABLE C.10: Baseline characteristics of OA Scoring HCA clusters with Ward linkage method

#	Variable	Cluster 1 (n=678)	Cluster 2 (n=341)	Cluster 3 (n=227)	Cluster 4 (n=214)	Cluster 5 (n=237)	Cluster 6 (n=91)	p-value
1	Lft_T0	55.94 ± 5.17	55.82 ± 5.14	55.66 ± 5.4	55.89 ± 5.36	56.51 ± 5.29	55.89 ± 5.12	0.70
2	RAS	99.12	96.48	97.36	95.79	98.31	96.7	NA
3	SEXE	79.5	78.59	77.97	80.84	80.17	69.23	NA
4	BMI	25.88 ± 3.85	26.33 ± 3.71	26.4 ± 4.22	25.66 ± 3.77	26.46 ± 4.49	26.23 ± 4.2	0.70
5	Menopauze_01	331 (75)	147 (76)	103 (75)	107 (79)	112 (81)	44 (86)	NA
6	Leptinengml	15.78 ± 14.69	18.01 ± 16.54	18.65 ± 17.81	16.1 ± 15.27	17.8 ± 17.53	17.49 ± 17.42	0.92
7	Adiponectineugml	12.13 ± 6.69	11.08 ± 6.53	10.5 ± 6.1	12.94 ± 6.85	12.78 ± 7.42	10.84 ± 5.23	0.49
8	Resistinengml	3.72 ± 1.25	3.83 ± 1.2	3.8 ± 1.31	3.76 ± 1.2	3.82 ± 1.48	3.74 ± 1.26	0.49
9	CTXIugmmol	180.36 ± 117.76	188.39 ± 131.59	169.28 ± 109.02	169.17 ± 100.08	179.9 ± 125.4	182.97 ± 118.58	0.91
10	uNTXInMBCEmmol	41.54 ± 21.01	41.91 ± 20.3	39.87 ± 18.04	42.2 ± 20.78	40.93 ± 20.26	39.95 ± 16.22	0.79
11	sPINP	45.87 ± 19.93	45.61 ± 19.38	44.51 ± 23.96	47.61 ± 18.11	46.77 ± 23.46	43.98 ± 14.96	0.77
12	sOC	14.28 ± 6.73	14.55 ± 7.17	13.79 ± 6.02	14.14 ± 5.05	14.7 ± 7.23	14.13 ± 6.44	0.54
13	sC12C	0.22 ± 0.37	0.19 ± 0.25	0.25 ± 0.54	0.25 ± 0.58	0.19 ± 0.09	0.25 ± 0.59	0.60
14	CTXIIngmmol	227.94 ± 132.23	226.15 ± 138.78	221.17 ± 106.78	215.8 ± 127.18	205.62 ± 115.39	218.09 ± 131.94	0.93
15	sCS846	78.41 ± 59.62	77.69 ± 46.09	79.02 ± 40.7	74.56 ± 28.44	79.39 ± 54.22	77.82 ± 24.87	0.29
16	sCOMPUl	8.8 ± 2.41	8.74 ± 2.26	8.54 ± 2.02	8.71 ± 2.34	8.76 ± 2.19	8.48 ± 2.31	0.62
17	sPIIANP	1483.74 ± 732.03	1577.71 ± 804.05	1439.19 ± 526.9	1539.5 ± 804.6	1549.26 ± 606.95	1465.24 ± 500.47	0.28
18	sHA	32.07 ± 23.64	34.03 ± 27.72	33.22 ± 32.01	33.16 ± 22.44	33.93 ± 27.28	37.95 ± 22.06	0.97
19	sPIIINP	4.27 ± 1.1	4.26 ± 1.32	4.47 ± 1.44	4.48 ± 1.2	4.24 ± 1.19	4.08 ± 1.08	0.86
20	hsCRP	2.49 ± 3.99	3.16 ± 4.26	3.99 ± 10.09	2.49 ± 4.72	3.27 ± 4.32	2.2 ± 2.6	0.78
21	BSE	9.46 ± 7.48	10.15 ± 7.53	10.58 ± 7.87	9.06 ± 6.37	10.17 ± 7.03	9.46 ± 7.1	0.92

Appendix D

Clustering Evaluation

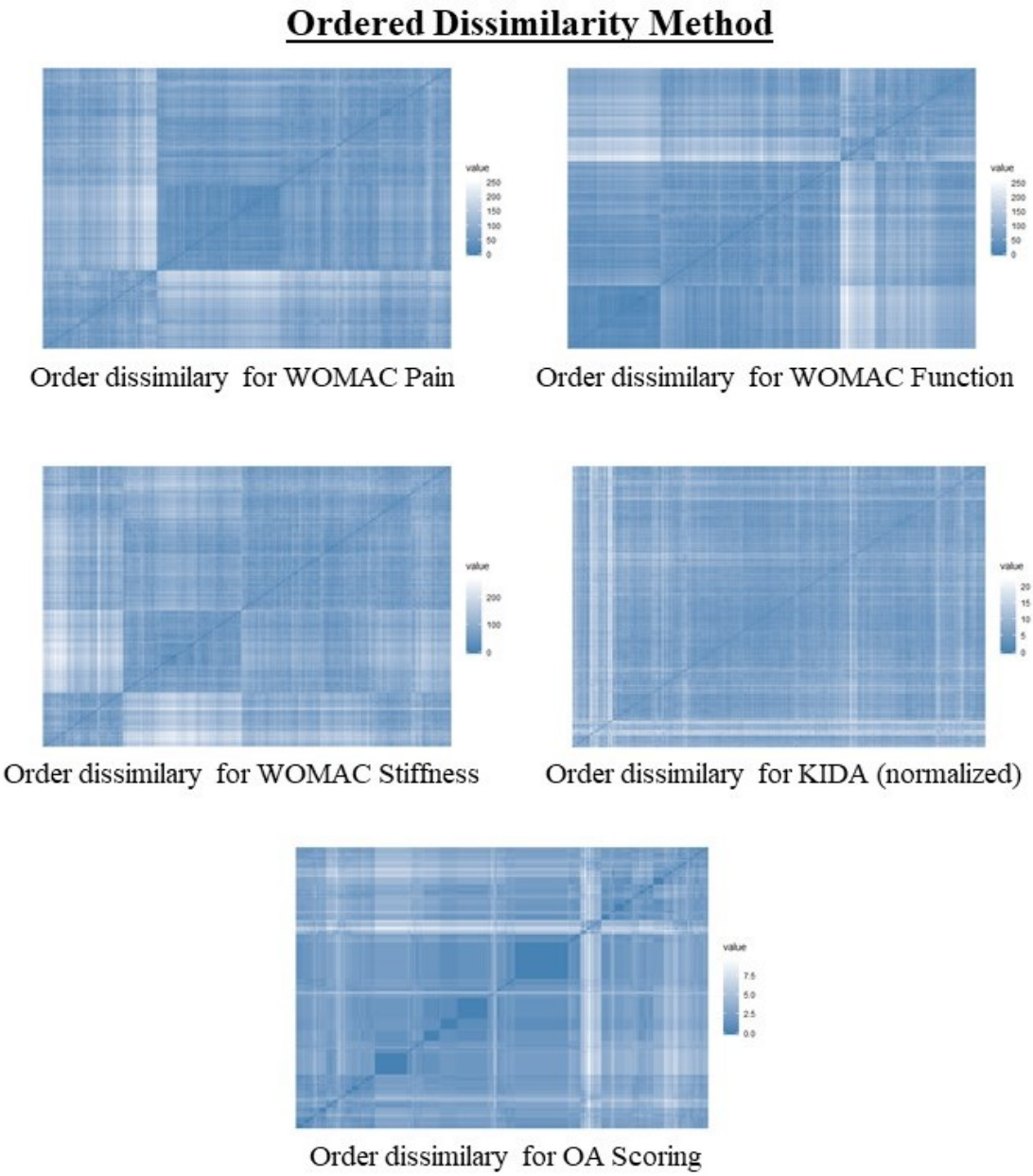


FIGURE D.1: Ordered Dissimilarity Plots.

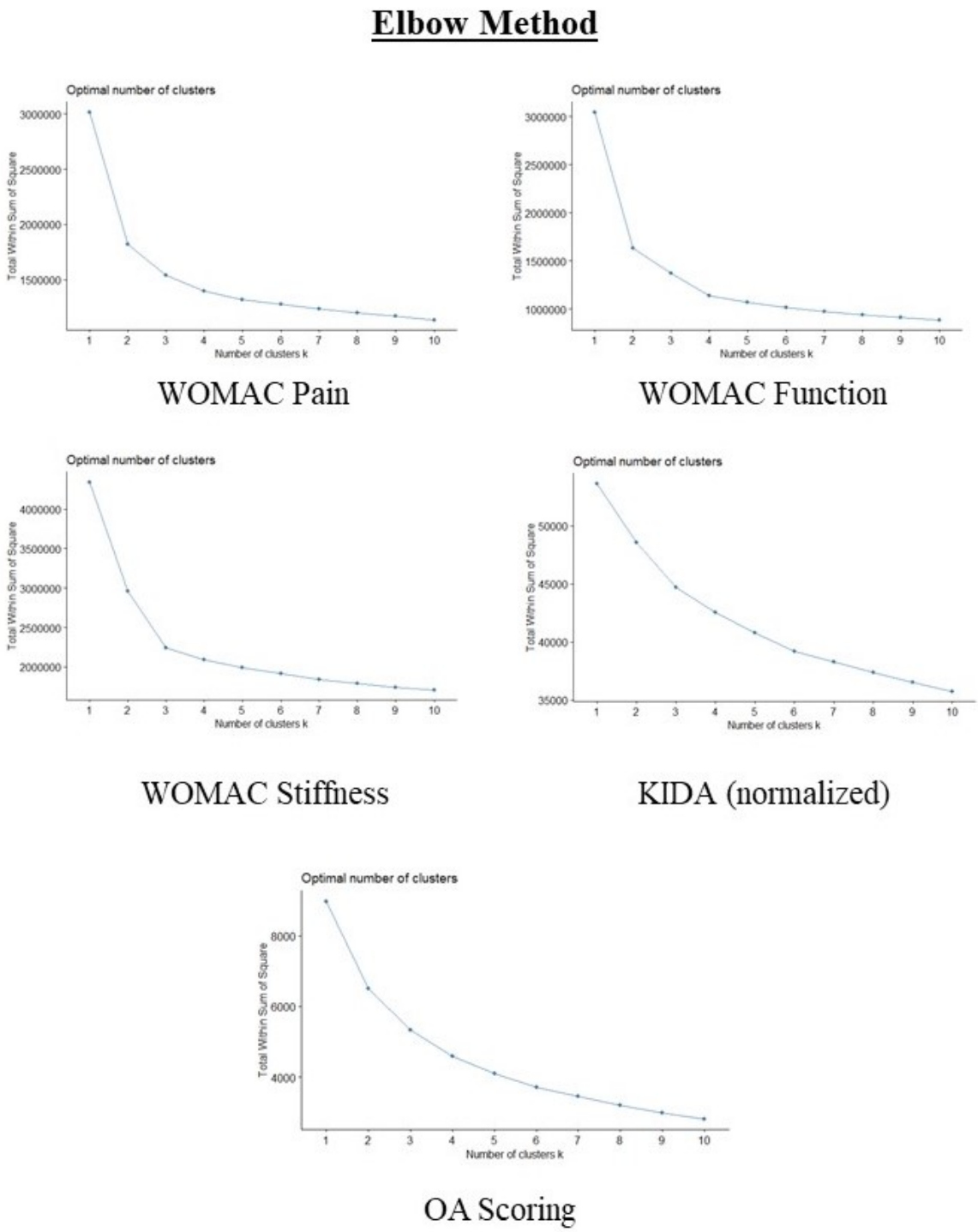


FIGURE D.2: Clustering evaluation with elbow method for HCA clusters.

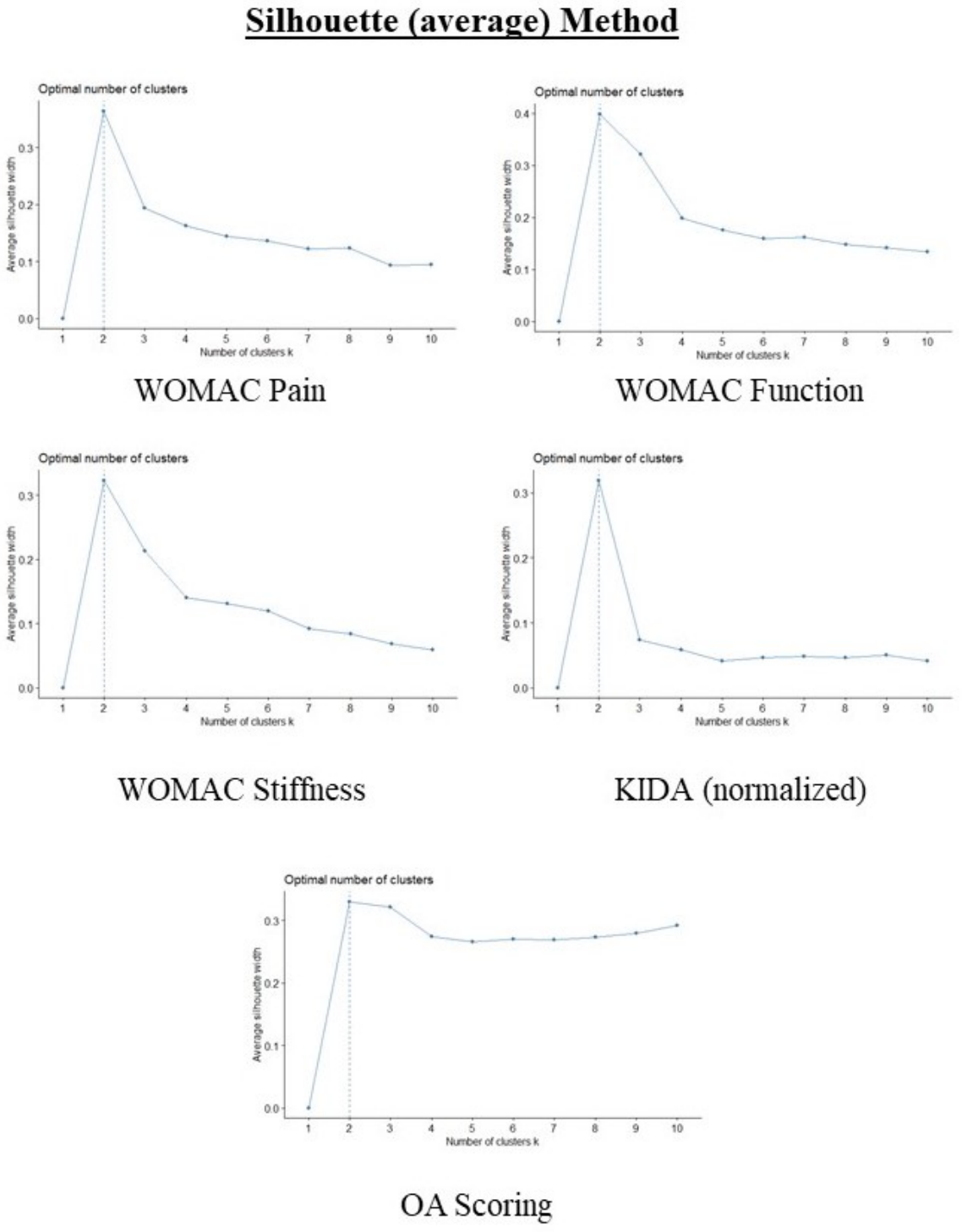


FIGURE D.3: Clustering evaluation with average silhouette coefficient for HCA clusters.

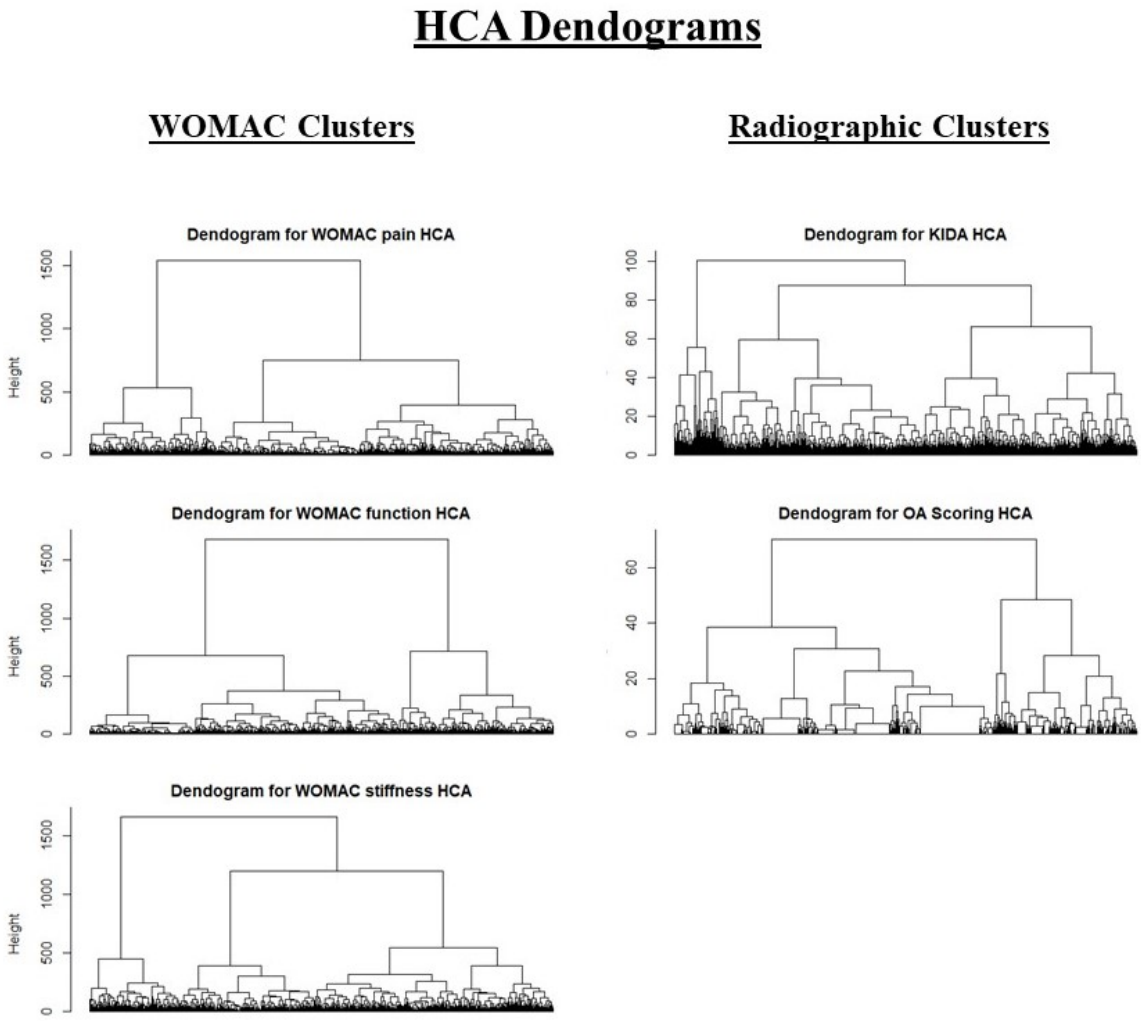


FIGURE D.4: Dendograms for HCA clusters.

TABLE D.1: Posterior probabilities of funHDDC clusters

	WOMAC Pain	**WOMAC Function**	**WOMAC Stiffness**	**KIDA**	**OA Scoring**
Cluster 1	0.93	0.96	0.92	0.94	0.99
Cluster 2	0.96	0.94	0.88	0.94	0.98
Cluster 3	0.96	0.92	0.86	0.98	1.00
Cluster 4	0.92	0.93	0.88	0.96	0.99
Cluster 5	0.91	0.92	0.93	0.97	1.00
Cluster 6				0.95	0.98
Cluster 7				0.95	
Cluster 8				0.95	
Average	*0.94*	*0.94*	*0.89*	*0.96*	*0.99*

Appendix E

Listings

```
K_SKY_OST_li_data <- CTidy %>%
  dplyr::select(year, nsin, K_SKY_OST_li) %>%
  as.data.frame()
# Casting the K_SKY_OST_li data and then selecting only year
columns (leaving out nsin)
recasted_K_SKY_OST_li <- cast(data = K_SKY_OST_li_data,
nsin~year, value = 'K_SKY_OST_li') %>%
  dplyr::select("0", '2', '5', '8', '10') %>% as.data.frame()
# NA imputation of K_SKY_OST_li
ind_K_SKY_OST_li_impute <-
which(rowSums(is.na(recasted_K_SKY_OST_li)) <= 3)
# Create new dataframe with <=3  NAs, which will be used for
# linear interpolation
X_K_SKY_OST_li_Max3NAs <-
as.data.frame(recasted_K_SKY_OST_li[ind_K_SKY_OST_li_impute,])
# Compute linear interpolation on the dataframe with at
# least 2 data points per row (Maximum 3 NAs)
X_K_SKY_OST_li_Max3NAs_inter <-
as.data.frame(na_interpolation((t(X_K_SKY_OST_li_Max3NAs)),
option = "linear"))
# Create index
ind_K_SKY_OST_li_inter <-
which(rowSums(is.na(t(X_K_SKY_OST_li_Max3NAs_inter))) == 0)
# Transpose
K_SKY_OST_li_inter <-
as.data.frame(t(X_K_SKY_OST_li_Max3NAs_inter))
# Create index by intersecting 2 variables
ind_radio_18 <- Reduce(base::intersect,
list(ind_KROsteophyteFemurLatmm_inter,
ind_K_SKY_OST_li_inter))
# Subset the variable with the index we just created
X_K_SKY_OST_li_inter <- K_SKY_OST_li_inter[ind_radio_18,]
# Create functional data object with previously created basis
# system
fd18_K_SKY_OST_li <- smooth.basis(argvals = c(0,2,5,8,10),
y = t(X_K_SKY_OST_li_inter), fdParobj = polyBasis)$fd
```

LISTING E.1: Example of code for specific data preparation for K_SKY_OST_li variable.

```
polyBasis <- create.polygonal.basis(argvals = c(0,2,5,8,10))
```

LISTING E.2: Example of code used for creating the polygonal basis.

```
res_poly_WOMAC <- funHDDC(list(fd1_wmpyns,fd2_wmfuns,
fd3_wmstfs), model=c("AkjBkQkDk","AkjBQkDk","AkBkQkDk",
"AkBQkDk","ABkQkDk","ABQkDk"), K=c(3:8))
```

LISTING E.3: Example of code used for running the funHDDC algorithm with three variables, testing all model variations, and testing a range of K from three to eight clusters.

```
# Variable: fd1_wmpyns, model: res_poly_WOMAC
select1 <-
fd(fd1_wmpyns$coefs[,which(res_poly_WOMAC$class==1)],
fd1_wmpyns$basis)
select2 <-
fd(fd1_wmpyns$coefs[,which(res_poly_WOMAC$class==2)],
fd1_wmpyns$basis)
select3 <-
fd(fd1_wmpyns$coefs[,which(res_poly_WOMAC$class==3)],
fd1_wmpyns$basis)
select4 <-
fd(fd1_wmpyns$coefs[,which(res_poly_WOMAC$class==4)],
fd1_wmpyns$basis)
select5 <-
fd(fd1_wmpyns$coefs[,which(res_poly_WOMAC$class==5)],
fd1_wmpyns$basis)
## Plot each of the curves
plot(mean.fd(select1),col="orange",ylim=c(0,100),lty=1,
lwd=3,ylab = "Mean Values",
main = "Standardized WOMAC Pain Scale")
lines(mean.fd(select2),col="palegreen2",lty=1,lwd=3)
lines(mean.fd(select3),col="navy",lty=1,lwd=3)
lines(mean.fd(select3),col="purple",lty=1,lwd=3)
lines(mean.fd(select3),col="red",lty=1,lwd=3)
```

LISTING E.4: Example of code used for plotting graphical representation of the groups mean curves for WOMAC Pain variable.

```
# Prepare the data as a matrix
HCA_wmpyns_data <- as.matrix(wmpyns_inter)
colnames(HCA_wmpyns_data) <- c("year_0", "year_1", "year_2",
"year_3", "year_4", "year_5", "year_6", "year_7", "year_8",
"year_9", "year_10")
# Calculate Euclidean distance
distance_matrix_wmpyns <- dist(HCA_wmpyns_data)
```

```
# Cluster dendrogram with desired linkage method
hc_wmpyns_w2 <- hclust(distance_matrix_wmpyns,
method = "ward.D2")
# Plot dendrogram
plot(hc_wmpyns_w2, hang = -1, main = "Dendrogram for WOMAC
pain HCA")
# Cut dendrogram tree at desired number of clusters K
hc_wmpyns_members <- cutree(hc_wmpyns_w2, 5)
# Create list of members per cluster
all_wmpyns_nsins <- cast(data = wmpyns_data, nsin~year,
value = 'wmpyns') %>% dplyr::select('nsin') %>% as.data.frame()
# Select sample with index: ind_wmpyns_inter
all_wmpyns_nsins <- all_wmpyns_nsins[ind_wmpyns_inter,]
# Create dataframe with subject ID and cluster membership
hca_members_clusters <- data.frame("nsin" = all_wmpyns_nsins,
"cluster" = hc_wmpyns_members)
# Create dataset with ID and clusters included
HCA_wmpyns_data_with_nsins <- as.data.frame(HCA_wmpyns_data)
HCA_wmpyns_data_with_nsins$nsins <- hca_members_clusters$nsin
HCA_wmpyns_data_with_nsins$clusters <-
hca_members_clusters$cluster
# Calculate cluster means
hca_wmpyns_cluster_means <-
aggregate(HCA_wmpyns_data_with_nsins,
list(hca_members_clusters$cluster), mean)
# Select only columns with means and transpose dataframe
hca_wmpyns_cluster_means <- t(hca_wmpyns_cluster_means[,1:12])
# Remove cluster number
hca_wmpyns_cluster_means <- hca_wmpyns_cluster_means[-1,]
# Plot cluster means
plot(hca_wmpyns_cluster_means[,1], col="orange",xlim=c(1,11),
ylim=c(0,100),lty=1,lwd=3, xlab = "time", ylab = "Mean Values",
main = "Standardized WOMAC Pain Scale", type = "l")
lines(hca_wmpyns_cluster_means[,2],col="palegreen2",lty=1,
lwd=3)
lines(hca_wmpyns_cluster_means[,3],col="navy",lty=1,lwd=3)
lines(hca_wmpyns_cluster_means[,4],col="purple",lty=1,lwd=3)
lines(hca_wmpyns_cluster_means[,5],col="red",lty=1,lwd=3)
grid(nx = NULL, ny = NULL,col = "lightgray", lty = "dotted")
lines(hca_wmpyns_cluster_means[,6],col="lightblue",lty=1,lwd=3)
lines(hca_wmpyns_cluster_means[,7],col="black",lty=1,lwd=3)
lines(hca_wmpyns_cluster_means[,8],col="pink",lty=1,lwd=3)
grid(nx = NULL, ny = NULL,col = "lightgray", lty = "dotted")
```

LISTING E.5: Example of code used for HCA for WOMAC Pain variable.

Bibliography

Abhishek, A. et al. (2016). "Does chondrocalcinosis associate with a distinct radiographic phenotype of osteoarthritis in knees and hips? A case–control study". In: *Arthritis care & research* 68.2, pp. 211–216.

Aggarwal C.and Reddy, C. (2014). *Data classification: algorithms and applications*. CRC press.

Aghabozorgi, S., A. Shirkhorshidi, and T. Wah (2015). "Time-series clustering–a decade review". In: *Information Systems* 53, pp. 16–38.

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Altman, R. and G. Gold (2007). "Atlas of individual radiographic features in osteoarthritis, revised". In: *Osteoarthritis and cartilage* 15, A1–A56.

Arden, N. and M. Nevitt (2006). "Osteoarthritis: epidemiology". In: *Best practice & research Clinical rheumatology* 20.1, pp. 3–25.

Attur, M. et al. (2011). "Increased interleukin-1β gene expression in peripheral blood leukocytes is associated with increased pain and predicts risk for progression of symptomatic knee osteoarthritis". In: *Arthritis & Rheumatism* 63.7, pp. 1908–1917.

Bedson, J. and P. Croft (2008). "The discordance between clinical and radiographic knee osteoarthritis: a systematic search and summary of the literature". In: *BMC musculoskeletal disorders* 9.1, p. 116.

Berlin, K. S, G. Parra, and N. Williams (2014). "An introduction to latent variable mixture modeling (part 2): longitudinal latent class growth analysis and growth mixture models". In: *Journal of Pediatric Psychology* 39.2, pp. 188–203.

Berry, P. et al. (2010a). "Markers of bone formation and resorption identify subgroups of patients with clinical knee osteoarthritis who have reduced rates of cartilage loss". In: *The Journal of rheumatology* 37.6, pp. 1252–1259.

Berry, P. et al. (2010b). "Relationship of serum markers of cartilage metabolism to imaging and clinical outcome measures of knee joint structure". In: *Annals of the rheumatic diseases* 69.10, pp. 1816–1822.

Bezdek, J. and R. Hathaway (2002). "VAT: A tool for visual assessment of (cluster) tendency". In: *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*. Vol. 3. IEEE, pp. 2225–2230.

Bieleman, H. et al. (2019). "Trajectories of Physical Work Capacity in Early Symptomatic Osteoarthritis of Hip and Knee: Results from the Cohort Hip and Cohort Knee (CHECK) Study". In: *Journal of occupational rehabilitation* 29.3, pp. 483–492.

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.

Bijlsma, J., F. Berenbaum, and F. Lafeber (2011). "Osteoarthritis: an update with relevance for clinical practice". In: *The Lancet* 377.9783, pp. 2115–2126.

Black, Ken (2019). *Business statistics: for contemporary decision making*. John Wiley & Sons.

Blashfield, R. and M. Aldenderfer (1988). "The methods and problems of cluster analysis". In: *Handbook of multivariate experimental psychology*. Springer, pp. 447–473.

Bouveyron, C., E. Côme, J. Jacques, et al. (2015). "The discriminative functional mixture model for a comparative analysis of bike sharing systems". In: *The Annals of Applied Statistics* 9.4, pp. 1726–1760.

Bouveyron, C., S. Girard, and C. Schmid (2007). "High-dimensional data clustering". In: *Computational Statistics & Data Analysis* 52.1, pp. 502–519.

Bouveyron, C. and J. Jacques (2011). "Model-based clustering of time series in group-specific functional subspaces". In: *Advances in Data Analysis and Classification* 5.4, pp. 281–300.

Bouveyron, C. et al. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Brown, T. et al. (2006). "Posttraumatic osteoarthritis: a first estimate of incidence, prevalence, and burden of disease". In: *Journal of orthopaedic trauma* 20.10, pp. 739–744.

Bruyere, O. et al. (2015). "Can we identify patients with high risk of osteoarthritis progression who will respond to treatment? A focus on epidemiology and phenotype of osteoarthritis". In: *Drugs & aging* 32.3, pp. 179–187.

Burnett, S. et al. (1994). *A radiographic atlas of osteoarthritis*.

Cardoso, J. et al. (2016). "Experimental pain phenotyping in community-dwelling individuals with knee osteoarthritis". In: *Pain* 157.9, p. 2104.

Carlesso, L. et al. (2019). "Pain susceptibility phenotypes in those free of knee pain with or at risk of knee osteoarthritis: the multicenter osteoarthritis study". In: *Arthritis & Rheumatology* 71.4, pp. 542–549.

Carlson, A. et al. (2019). "Characterization of synovial fluid metabolomic phenotypes of cartilage morphological changes associated with osteoarthritis". In: *Osteoarthritis and cartilage* 27.8, pp. 1174–1184.

Castañeda, S. et al. (2014). *Osteoarthritis: a progressive disease with changing phenotypes*.

Chapman, P. et al. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. Tech. rep. The CRISP-DM consortium.

Conaghan, P. et al. (2010). "Clinical and ultrasonographic predictors of joint replacement for knee osteoarthritis: results from a large, 3-year, prospective EULAR study". In: *Annals of the rheumatic diseases* 69.4, pp. 644–647.

Cotofana S.and Wyman, Bradley T et al. (2013). "Relationship between knee pain and the presence, location, size and phenotype of femorotibial denuded areas of subchondral bone as visualized by MRI". In: *Osteoarthritis and cartilage* 21.9, pp. 1214–1222.

Cross, M. et al. (2014). "The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study". In: *Annals of the rheumatic diseases* 73.7, pp. 1323–1330.

Cruz-Almeida, Y. et al. (2013). "Psychological profiles and pain characteristics of older adults with knee osteoarthritis". In: *Arthritis care & research* 65.11, pp. 1786–1794.

Delaigle, A., P. Hall, et al. (2010). "Defining probability density for a distribution of random functions". In: *The Annals of Statistics* 38.2, pp. 1171–1193.

Dell'Isola, A. et al. (2016). "Identification of clinical phenotypes in knee osteoarthritis: a systematic review of the literature". In: *BMC musculoskeletal disorders* 17.1, p. 425.

Deveza, L. and R. Loeser (2018). "Is osteoarthritis one disease or a collection of many?" In: *Rheumatology* 57.suppl_4, pp. iv34–iv42.

Deveza, L., A. Nelson, and R. Loeser (2019). "Phenotypes of osteoarthritis-current state and future implications". In: *Clinical and experimental rheumatology* 37.Suppl 120, p. 64.

Deveza, L. et al. (2017). "Knee osteoarthritis phenotypes and their relevance for outcomes: a systematic review". In: *Osteoarthritis and cartilage* 25.12, pp. 1926–1941.

Deveza, L. et al. (2019). "Trajectories of femorotibial cartilage thickness among persons with or at risk of knee osteoarthritis: development of a prediction model to identify progressors". In: *Osteoarthritis and cartilage* 27.2, pp. 257–265.

Driban, J. et al. (2010). "Is osteoarthritis a heterogeneous disease that can be stratified into subsets?" In: *Clinical rheumatology* 29.2, p. 123.

Driver, H. and A. Kroeber (1932). *Quantitative expression of cultural relationships*. Vol. 31. 4. University of California Press.

Egsgaard, L. et al. (2015). "Identifying specific profiles in patients with different degrees of painful knee osteoarthritis based on serological biochemical and mechanistic pain biomarkers: a diagnostic approach based on cluster analysis". In: *Pain* 156.1, pp. 96–107.

Elbaz, A. et al. (2014). "Novel classification of knee osteoarthritis severity based on spatiotemporal gait analysis". In: *Osteoarthritis and cartilage* 22.3, pp. 457–463.

Ferraty, Frédéric and Philippe Vieu (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Fox, Wendy R (1991). "Finding groups in data: an introduction to cluster analysis". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 40.3, pp. 486–487.

Frey-Law L.and Bohr, Nicole L et al. (2016). "Pain sensitivity profiles in patients with advanced knee osteoarthritis". In: *Pain* 157.9, p. 1988.

Geenen, R. and J. Bijlsma (2010). "Psychological management of osteoarthritic pain". In: *Osteoarthritis and cartilage* 18.7, pp. 873–875.

Glicksberg, B. et al. (2019). "Elucidating genetic associations that differentiate pain progression cluster groups in knee osteoarthritis patients". In: *Osteoarthritis and Cartilage* 27, S294–S295.

Han, Jiawei, Jian Pei, and Micheline Kamber (2011). *Data mining: concepts and techniques*. Elsevier.

Happ-Kurz, C. (2020). "Object-Oriented Software for Functional Data". In: *Journal of Statistical Software* 93.5, pp. 1–38.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Heard, B. et al. (2013). "Intraarticular and systemic inflammatory profiles may identify patients with osteoarthritis". In: *The Journal of rheumatology* 40.8, pp. 1379–1387.

Hoogeboom, T. et al. (2012). "Joint-pain comorbidity, health status, and medication use in hip and knee osteoarthritis: A cross-sectional study". In: *Arthritis care & research* 64.1, pp. 54–58.

Hopkins, Brian and John Gordon Skellam (1954). "A new method for determining the type of distribution of plant individuals". In: *Annals of Botany* 18.2, pp. 213–227.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of classification* 2.1, pp. 193–218.

Ieva, F. et al. (2013). "Multivariate functional clustering for the morphological analysis of electrocardiograph curves". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62.3, pp. 401–418.

Iijima, H. et al. (2015). "Clinical phenotype classifications based on static varus alignment and varus thrust in Japanese patients with medial knee osteoarthritis". In: *Arthritis & rheumatology* 67.9, pp. 2354–2362.

Jacques, J. and C. Preda (2014a). "Functional data clustering: a survey". In: *Advances in Data Analysis and Classification* 8.3, pp. 231–255.

— (2014b). "Model-based clustering for multivariate functional data". In: *Computational Statistics & Data Analysis* 71, pp. 92–106.

Jacques, Julien and Cristian Preda (2013). "Funclust: A curves clustering method using functional random variables density approximation". In: *Neurocomputing* 112, pp. 164–171.

James, G. and C. Sugar (2003). "Clustering for sparsely sampled functional data". In: *Journal of the American Statistical Association* 98.462, pp. 397–408.

James, G. et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Jenkins, J. and Thomas P McCoy (2015). "Symptom clusters, functional status, and quality of life in older adults with osteoarthritis". In: *Orthopaedic Nursing* 34.1, pp. 36–42.

Johannsen, W. (1911). "The genotype conception of heredity". In: *The American Naturalist* 45.531, pp. 129–159.

Karsdal, M. et al. (2015). "OA phenotypes, rather than disease stage, drive structural progression–identification of structural progressors from 2 phase III randomized clinical studies with symptomatic knee OA". In: *Osteoarthritis and cartilage* 23.4, pp. 550–558.

Karsdal, M. et al. (2016). "Disease-modifying treatments for osteoarthritis (DMOADs) of the knee and hip: lessons learned from failures and opportunities for the future". In: *Osteoarthritis and Cartilage* 24.12, pp. 2013–2021.

Karsdal, M. et al. (2019). "Serological biomarker profiles of rapidly progressive osteoarthritis in tanezumab-treated patients". In: *Osteoarthritis and cartilage* 27.3, pp. 484–492.

Kassambara, A. and F. Mundt (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7. URL: https://CRAN.R-project.org/package=factoextra.

Kaufman, Leonard and Peter J Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Kellgren, J. and J. Lawrence (1957). "Radiological assessment of osteo-arthrosis". In: *Annals of the rheumatic diseases* 16.4, p. 494.

Kittelson, A., Jennifer E Stevens-Lapsley, and Sarah J Schmiege (2016). "Determination of pain phenotypes in knee osteoarthritis: a latent class analysis using data from the osteoarthritis initiative". In: *Arthritis care & research* 68.5, pp. 612–620.

Knoop, J. et al. (2011). "Identification of phenotypes with different clinical outcomes in knee osteoarthritis: data from the Osteoarthritis Initiative". In: *Arthritis care & research* 63.11, pp. 1535–1542.

Kotlarz, H. et al. (2010). "Osteoarthritis and absenteeism costs: evidence from US National Survey Data". In: *Journal of occupational and environmental medicine* 52.3, pp. 263–268.

Kovac, S. et al. (2008). "Association of health-related quality of life with dual use of prescription and over-the-counter nonsteroidal antiinflammatory drugs". In: *Arthritis Care & Research* 59.2, pp. 227–233.

Lee, A. et al. (2018). "Pain and functional trajectories in symptomatic knee osteoarthritis over up to 12 weeks of exercise exposure". In: *Osteoarthritis and cartilage* 26.4, pp. 501–512.

Lee, S. et al. (2015). "Obesity, metabolic abnormality, and knee osteoarthritis: a cross-sectional study in Korean women". In: *Modern rheumatology* 25.2, pp. 292–297.

Losina, E. et al. (2013). "Lifetime risk and age at diagnosis of symptomatic knee osteoarthritis in the US". In: *Arthritis care & research* 65.5, pp. 703–711.

Loza, E. et al. (2009). "Economic burden of knee and hip osteoarthritis in Spain". In: *Arthritis Care & Research* 61.2, pp. 158–165.

Marijnissen, A. et al. (2008). "Knee Images Digital Analysis (KIDA): a novel method to quantify individual radiographic features of knee osteoarthritis in detail". In: *Osteoarthritis and cartilage* 16.2, pp. 234–243.

Menger, V. et al. (2016). "Transitioning to a Data Driven Mental Health Practice: Collaborative Expert Sessions for Knowledge and Hypothesis Finding". In: *Computational and Mathematical Methods in Medicine* 2016, pp. 1–12.

Meulenbelt, I. et al. (2007). "Clusters of biochemical markers are associated with radiographic subtypes of osteoarthritis (OA) in subject with familial OA at multiple sites. The GARP study". In: *Osteoarthritis and cartilage* 15.4, pp. 379–385.

Moher, D. et al. (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement". In: *Annals of internal medicine* 151.4, pp. 264–269.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018). *Foundations of Machine Learning*. 2nd. The MIT Press. ISBN: 0262039400.

Moritz, S. and T. Bartz-Beielstein (2017). "imputeTS: Time Series Missing Value Imputation in R". In: *The R Journal* 9.1, pp. 207–218.

Munugoda, Ishanka P et al. (2020). "Identifying subgroups of community-dwelling older adults and their prospective associations with long-term knee osteoarthritis outcomes". In: *Clinical Rheumatology*, pp. 1–9.

Murphy, S. et al. (2011). "Subgroups of older adults with osteoarthritis based upon differing comorbid symptom presentations and potential underlying pain mechanisms". In: *Arthritis research & therapy* 13.4, R135.

Nelson, A. et al. (2013). "Brief report: differences in multijoint symptomatic osteoarthritis phenotypes by race and sex: the Johnston County Osteoarthritis Project". In: *Arthritis & Rheumatism* 65.2, pp. 373–377.

Nelson, A. et al. (2019). "A machine learning approach to knee osteoarthritis phenotyping: data from the FNIH Biomarkers Consortium". In: *Osteoarthritis and cartilage* 27.7, pp. 994–1001.

Osgood, E. et al. (2015). "Development of a bedside pain assessment kit for the classification of patients with osteoarthritis". In: *Rheumatology international* 35.6, pp. 1005–1013.

Osteoarthritis (2019). URL: https://www.volksgezondheidenzorg.info/onderwerp/artrose/cijfers-context/oorzaken-en-gevolgen#!node-risicofactoren-en-gevolgen-van-artrose.

Palazzo, C. et al. (2016). "Risk factors and burden of osteoarthritis". In: *Annals of physical and rehabilitation medicine* 59.3, pp. 134–138.

Pan, F. et al. (2019). "Differentiating knee pain phenotypes in older adults: a prospective cohort study". In: *Rheumatology* 58.2, pp. 274–283.

Pan, F. et al. (2020). "Metabolic syndrome and trajectory of knee pain in older adults". In: *Osteoarthritis and Cartilage* 28.1, pp. 45–52.

Peat, G. et al. (2012). "Clinical features of symptomatic patellofemoral joint osteoarthritis". In: *Arthritis research & therapy* 14.2, R63.

Pinto, L. et al. (2015). "Derivation and validation of clinical phenotypes for COPD: a systematic review". In: *Respiratory research* 16.1, p. 50.

Prieto-Alhambra, D. et al. (2014). "Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints". In: *Annals of the rheumatic diseases* 73.9, pp. 1659–1664.

Qannari, El Mostafa, Philippe Courcoux, and Pauline Faye (2014). "Significance test of the adjusted Rand index. Application to the free sorting task". In: *Food quality and preference* 32, pp. 93–97.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

— (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramsay, J. (1982). "When the data are functions". In: *Psychometrika* 47.4, pp. 379–396.

Ramsay, J.., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. 1st. Springer Publishing Company, Incorporated.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. 2nd. Springer Science Business Media, Inc.

Ramsay, J. O. et al. (2018). *fda: Functional Data Analysis*. R package version 2.4.8. URL: https://CRAN.R-project.org/package=fda.

Rathbun A.and Schuler, Megan S et al. (2020). "Depression subtypes in persons with or at risk for symptomatic knee osteoarthritis". In: *Arthritis care & research* 72.5, p. 669.

Roze, R. et al. (2016). "Differences in MRI features between two different osteoarthritis subpopulations: data from the Osteoarthritis Initiative". In: *Osteoarthritis and cartilage* 24.5, pp. 822–826.

Runhaar, J. et al. (2018). *SAT0557 10-year trajectories of pain in early knee and hip osteoarthritis; the check study*.

Schiphof, D. et al. (2019). "The clinical and radiographic course of early knee and hip osteoarthritis over 10 years in CHECK (Cohort Hip and Cohort Knee)". In: *Osteoarthritis and cartilage* 27.10, pp. 1491–1500.

Schmutz, A., J. Jacques, and C. Bouveyron (2019). *funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces*. R package. URL: https://cran.r-project.org/web/packages/funHDDC/index.html.

Schmutz, A. et al. (2020). "Clustering multivariate functional data in group-specific functional subspaces". In: *Computational Statistics*, pp. 1–31.

Schwarz, Gideon et al. (1978). "Estimating the dimension of a model". In: *The annals of statistics* 6.2, pp. 461–464.

Sharif, M. et al. (Nov. 2006). "Serum cartilage oligomeric matrix protein and other biomarker profiles in tibiofemoral and patellofemoral osteoarthritis of the knee". In: *Rheumatology* 45.5, pp. 522–526.

Siebuhr, A. et al. (2014). "Identification and characterisation of osteoarthritis patients with inflammation derived tissue turnover". In: *Osteoarthritis and cartilage* 22.1, pp. 44–50.

Singhal, A. and D. Seborg (2005). "Clustering multivariate time-series data". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 19.8, pp. 427–438.

Snyder, H. (2019). "Literature review as a research methodology: An overview and guidelines". In: *Journal of Business Research* 104, pp. 333–339.

Soul, J. et al. (2018). "Stratification of knee osteoarthritis: two major patient subgroups identified by genome-wide expression analysis of articular cartilage". In: *Annals of the rheumatic diseases* 77.3, pp. 423–423.

Sowers, M. et al. (2009). "Knee osteoarthritis in obese women with cardiometabolic clustering". In: *Arthritis Care & Research* 61.10, pp. 1328–1336.

Tan, P., M. Steinbach, and V. Karpatne A.and Kumar (2019). *Introduction to Data Mining*. 2nd. Pearson.

Tellaroli, P. et al. (2018). *CrossClustering: A Partial Clustering Algorithm*. R package version 4.0.3. URL: https://CRAN.R-project.org/package=CrossClustering.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie (2001). "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411–423.

Tokushige, S., H. Yadohisa, and K. Inada (2007). "Crisp and fuzzy k-means clustering algorithms for multivariate functional data". In: *Computational Statistics* 22.1, pp. 1–16.

Tonge, D., M. Pearson, and S. Jones (2014). "The hallmarks of osteoarthritis and the potential to develop personalised disease-modifying pharmacological therapeutics". In: *Osteoarthritis and cartilage* 22.5, pp. 609–621.

Törmälehto, S. et al. (2019). "Eight-year trajectories of changes in health-related quality of life in knee osteoarthritis: Data from the Osteoarthritis Initiative (OAI)". In: *PloS one* 14.7.

United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Population Prospects 2019: Highlights (ST/ESA/SER.A/423)*. URL: https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf.

Van De Weerd, I. and S. Brinkkemper (2009). "Meta-modeling for situational analysis and design methods". In: *Handbook of research on modern systems analysis and design technologies and applications*. IGI Global, pp. 35–54.

Van Spil, W.E. (2012). "CHECKing biochemical markers in early-stage knee and hip osteoarthritis, a critical appraisal". PhD thesis. Utrecht University.

Van Spil, W.E. et al. (2010). "Serum and urinary biochemical markers for knee and hip-osteoarthritis: a systematic review applying the consensus BIPED criteria". In: *Osteoarthritis and cartilage* 18.5, pp. 605–612.

Van Spil, W.E. et al. (2012). "Clusters within a wide spectrum of biochemical markers for osteoarthritis: data from CHECK, a large cohort of individuals with very early symptomatic osteoarthritis". In: *Osteoarthritis and cartilage* 20.7, pp. 745–754.

Van Spil, W.E. et al. (2020). "A consensus-based framework for conducting and reporting osteoarthritis phenotype research". In: *Arthritis research & therapy* 22.1, pp. 1–7.

Vermunt, J. and J. Magidson (2002). "Latent class cluster analysis". In: *Applied latent class analysis* 11.89-106, p. 60.

Vongsirinavarat, M. et al. (2020). "Identification of knee osteoarthritis disability phenotypes regarding activity limitation: a cluster analysis". In: *BMC Musculoskeletal Disorders* 21.1, pp. 1–8.

Vos, T. et al. (2017). "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016". In: *The Lancet* 390.10100, pp. 1211–1259.

Waarsing, J., S. Bierma-Zeinstra, and H. Weinans (2015). "Distinct subtypes of knee osteoarthritis: data from the Osteoarthritis Initiative". In: *Rheumatology* 54.9, pp. 1650–1658.

Wang, H. et al. (2002). "Clustering by pattern similarity in large data sets". In: *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pp. 394–405.

Wang, J., J. Chiou, and H. Müller (2016). "Review of Functional Data Analysis". In: *Annual Review of Statistics and Its Application* 3, pp. 257–295.

Ward, J. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58, pp. 236–244.

Wesseling, J. et al. (2009). "CHECK (Cohort Hip and Cohort Knee): similarities and differences with the Osteoarthritis Initiative". In: *Annals of the rheumatic diseases* 68.9, pp. 1413–1419.

Wesseling, J. et al. (2014). "Cohort profile: cohort hip and cohort knee (CHECK) study". In: *International journal of epidemiology* 45.1, pp. 36–44.

White, D. et al. (2010). "Do worsening knee radiographs mean greater chances of severe functional limitation?" In: *Arthritis care & research* 62.10, pp. 1433–1439.

Whittle, R. et al. (2016). "Average symptom trajectories following incident radiographic knee osteoarthritis: data from the Osteoarthritis Initiative". In: *RMD open* 2.2.

Wickham, H. (2007). "Reshaping data with the reshape package". In: *Journal of Statistical Software* 21.12. URL: http://www.jstatsoft.org/v21/i12/paper.

Wickham, H. et al. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.4. URL: https://CRAN.R-project.org/package=dplyr.

Wieringa, R. (2014). *Design science methodology for information systems and software engineering*. Springer.

Wohlin, C. (2014). "Guidelines for snowballing in systematic literature studies and a replication in software engineering". In: *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pp. 1–10.

Wohlin, C. et al. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Wood, M. et al. (2019). "Macrophage proliferation distinguishes 2 subgroups of knee osteoarthritis patients". In: *JCI insight* 4.2.

Yamamoto, M. (2012). "Clustering of functional data in a low-dimensional subspace". In: *Advances in Data Analysis and Classification* 6.3, pp. 219–247.

Yao, F., H. Müller, and J. Wang (2005). "Functional data analysis for sparse longitudinal data". In: *Journal of the American Statistical Association* 100.470, pp. 577–590.

Young-Shand, K. et al. (2020). "Characterization of Total Knee Arthroplasty Patient: Clinical and Biomechanical Variability by Cluster Analysis". In: *Orthopaedic Proceedings*. Vol. 102. SUPP_1. The British Editorial Society of Bone & Joint Surgery, pp. 141–141.

Zambom, A., J. Collazos, and R. Dias (2019). "Functional data clustering via hypothesis testing k-means". In: *Computational Statistics* 34.2, pp. 527–549.

Zhang, S., C. Zhang, and Q. Yang (2003). "Data preparation for data mining". In: *Applied artificial intelligence* 17.5-6, pp. 375–381.

Zhang, W. et al. (2014). "Classification of osteoarthritis phenotypes by metabolomics analysis". In: *BMJ open* 4.11.

Zhao, K. et al. (2018). "Diagnosis of Osteoarthritis Subtypes with Blood Biomarkers". In: *bioRxiv*, p. 366047.