Leiden University
Statistics and Data Science

Combination of cluster analyses in longitudinal data from patients with Early Rheumatoid Arthritis

Supervisor:
Dr. Anikó Lovik

Student: **Gabriele Fresia**
Student number: **3121283**

February 2024

Contents

List of Abbreviations

ACPA	Anti-cyclic citrullinated peptide antibody
CareRA	Care in Early Rheumatoid Arthritis
CART	Classification and regression trees
CF	Clinical factor
CRP	C-reactive protein
DAS	Disease Activity Score
DAS28	Disease Activity Score measuring only 28 joints
EFA	Exploratory factor analysis
ESR	Erythrocyte sedimentation rate
HAQ	Health Assessment Questionnaire
HR	High-risk
LF	Laboratory factor
LR	Low-risk
MAR	Missing at random
MI	Multiple imputation
MICE	Multiple imputation by chained equations
MO	Multiple outputation
PaGH	Patient global health assessment
PCA	Principal component analysis
PhGH	Physician global health assessment
PRF	Patient-reported factor
QoL	Quality of Life
RA	Rheumatoid arthritis

RAQoL	Rheumatoid Arthritis Quality of Life questionnaire
RF	Rheumatoid factor
SD	Standard deviation
SJC	Swollen joint count
SJC28	28 swollen joint count
T2T	Treat-to-target
TJC	Tender joint count
TJC28	28 tender joint count
VAS	Visual analog scale
W8	Week 8

Summary

Background. Rheumatoid arthritis (RA) is a chronic, autoimmune disease affecting around 1% of the general population. Despite the availability of treatments aimed at controlling the disease, some patients still report unmet needs such as functional disability and fatigue.

Motivating case study. Care in Early Rheumatoid Arthritis (CareRA) was a 2-year clinical trial conducted in Belgium. 379 newly-diagnosed patients were assessed over ten visits.

Objective. To find clinically meaningful clusters of patients characterized by common socio-demographic characteristics and comparable disease activity.

Methods. The variables considered for the analyses consisted of clinical and laboratory measurements of disease activity and patient-reported outcomes such as pain and fatigue. Our study comprised two primary analyses: one cross-sectional and one using longitudinal data. In the former, data from Baseline and Week 8 were analyzed separately. Routinely collected clinical variables were recoded into three factors. Hierarchical K-means clustering was carried out independently at each visit. The study on longitudinal data was designed to be a proof-of-concept and covered the other eight visits. 2000 datasets were obtained by carrying out multiple imputation (20X) and within-cluster resampling (100X) to deal with incompleteness and correlated data. Hierarchical K-means clustering was conducted on each of the datasets, separately. Clusters were aligned using the centroids produced by the cluster analysis applied to the centroids of all datasets as the "gold standard".

Results. Both analyses uncovered the existence of four clusters. At Baseline, a specific subgroup of the RA population characterized by higher mean age and a higher proportion of men was identified as vulnerable. In the longitudinal analysis, the largest subset of patients comprised those who responded positively to the treatment and reported a relatively low disease burden. Other clusters included patients with unmet psychological needs and patients with chronic inflammation.

Interpretation and clinical relevance. Clusters were labelled and interpreted in collaboration with a team of clinical experts, who also provided the data. Patients with a rapid and sustained response to treatment could be referred to a family practitioner. A more holistic, patient-specific medical strategy could be developed to control the disease and the psychological distress of those patients for whom the disease burden is still high.

Conclusions. Through cluster analysis, it is possible to identify patients with a shared disease activity pattern and similar patient-reported outcomes which allows practitioners to employ individual treatment plans. The study provides a blueprint of the workflow, and decisions necessary for applying the methodology to a wider, more heterogeneous population.

Keywords: CareRA study, clinical trial, cluster analysis, exploratory factor analysis, multiple imputation, multiple outputation, rheumatoid arthritis

Chapter 1

Introduction

1.1 An introduction to rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic, autoimmune disease. The main clinical manifestation is the inflammation of the peripheral joints. It initially affects the small joints, then spreads to the larger joints, until it eventually affects the heart, lungs, and skin [1]. The expression and the development of the disease are highly heterogeneous among individuals. The most common symptoms are stiffness, swelling, and pain [2].

A study from 2021 reported the lifetime global prevalence of RA to be around 1% of the general population. There is a higher prevalence of RA in women than in men, with a ratio of 3:1, although in some geographical areas the proportion reaches a peak of 5:1 [3]. The median age of onset for the disease is 58 years. Nonetheless, the range of age at onset is very broad, with patients being diagnosed as early as 18 years old [4]. The incidence rate grows with age and is highest for the population above 65 years old, to quickly decline in the oldest-old [5]. Variability is exhibited as well in the disease prevalence across different populations. North American and Northern European areas estimate a worldwide highest prevalence of 0.5 – 1.1%. The lowest levels are registered in developing countries, which report a prevalence of the disease between 0.1% and 0.5% of the general population [6]. Both genetic and environmental factors could play a role in this discrepancy. Furthermore, the medical costs for RA patients are substantial, which could lead to an under-reporting in developing countries [7].

Patients with RA have around 50% increased chance of premature mortality, leading to a life expectancy that is from 3 to 10 years lower when compared to the general population, depending on the geographical area [8]. RA is rare in men aged 80 years or more. Excess premature mortality is likely the cause for the low prevalence of RA in the oldest-old [5]. A third to half of the premature mortality in RA patients is caused by an increase in comorbid cardiovascular diseases [9]. Other causes of death in patients with RA are infections and renal diseases. Some studies also report an excess in mortality from respiratory diseases [10].

Even when it does not lead to premature death, RA has a substantial impact on the daily activities and the quality of life of the patients [11]. The consequences of the physical disability caused by RA may constrain the patient's ability to execute basic daily activities such as dressing, walking, and writing. This reduction in physical ability has been found to be associated with fatigue, pain, and depression [12]. Moreover, depression alters the quality of life of the patients beyond the mental illness itself. When evaluating the course of the disease, patients who are also affected by comorbid depression have worse RA-related

outcomes. Rheumatologists rarely used to communicate about the burden of depression with RA patients affected by moderately severe and severe depression. As a consequence, the impact of depression on the disease endured and was not addressed sufficiently by the health regimens [13]. This is usually referred to as "unmet needs".

In general, the mental health burden is not the only affliction of patients with RA. The range of complaints may vary from functional impairment to constitutional manifestation, such as loss of general health or fatigue [14]. In RA, the variability in symptoms, as well as their fluctuations over time, may interfere with the diagnosis and the evaluation of the disease status. Moreover, when dealing with a chronic disease such as RA, rheumatologists are interested in evaluating the disease course, also known as disease activity, over time.

For decades, the course of RA disease activity has been evaluated through a great number of variables, since its severity cannot be assessed through a single variable [14]. To be able to measure the disease activity in daily clinical practice, the Disease Activity Score (DAS) was developed. The DAS integrates various pieces of information, such as the presence of swollen and tender joints or the patient's self-reported overall health, into a single continuous measure [15], [16]. In comparison to selecting multiple disease activity variables, the DAS has proved to have higher "criterion, correlation, and construct validity" [17].

A simplified version of the disease activity evaluation was then developed, measuring only 28 joints in the joint count. This examination, named DAS28, does not appear to produce a loss of meaningful information when compared to the traditional procedure [18]. Some of the laboratory variables included in the DAS and the DAS28 are the erythrocyte sedimentation rate and the C-reactive protein, biomarkers that help identify an inflammation. The clinical examination comprises the evaluation of tender joints and swollen joints. Lastly, both physician-assessed and self-reported global health are included.

Through the application of thresholds, the measures were used to differentiate between remission, and low, moderate, and high disease activity [14]. Remission is a state of absence of disease activity or a stage where it is so low that it is not burdensome to the patient [19]. As already mentioned, active disease is associated with physical disability and results in a reduction in quality of life and premature mortality [20]. Hence, a treat-to-target (T2T) strategy is recommended for patients with RA. A T2T approach is a medical strategy with remission as the ultimate goal. Through shared decision-making between the rheumatologist and the patient, disease management objectives are defined. The goal is to improve the prognosis of the disease. Accordingly, the disease activity is measured periodically to evaluate the specific strategy. The medications and their dosage can be modified, according to specific guidelines, in order to achieve the goal [21].

However, even patients in remission, or with low disease activity under T2T strategies, report residual symptoms [14]. Some of the most reported residual symptoms are functional disability, pain, impairment in quality of life, fatigue, mental health, sleep disturbance, morning stiffness, and decrease in work productivity [11]. When patients were asked to identify the most important factors in defining remission, the most common domains were the lack or reduction of pain and fatigue and the progress or preservation of independence [22].

1.2 Clinical problem

The residual symptoms and the inability of disease activity measures to capture the patient's burden led to a paradigm shift in rheumatology. Rheumatologists started hypothesizing that incorporating the patient-reported outcomes would be beneficial in daily clinical practice, clinical trials, as well as long-term observational studies. Pazmino et al. combined pain, fatigue, and physical function with the standard measurements to evaluate the disease activity [14]. They hypothesized that including these outcomes would contribute to a further understanding of the disease experience, separate from the traditional measures. Since many of the variables in the model had similar patterns of response, with subgroups of variables having high correlation, the interest laid in detecting latent, unobservable variables that defined the response.

Through exploratory factor analysis, the common factors underlying the observed variables were identified. When considering only the variables used in the standard procedure, a 2-factor model proved to be optimal. The two factors were the clinical assessment and the laboratory assessment. In contrast, when including the patient-reported outcomes, a three-factor model proved to best represent the disease status. On top of the previously mentioned factors, a third factor was uncovered denoting the patient-reported outcomes.

The study revealed the importance of including patient-reported outcomes in predicting unmet patient needs. Patients who are in sustained remission based on the treating rheumatologist's assessment may still suffer from the psychological burden of the disease [23]. Since disease activity only partly explains the burden of the disease, identifying groups of patients who may need additional help (for example, through referral to a mental health specialist) is an important clinical problem.

1.3 Methodological problem

The identification of a grouping of patients that is time-independent is a methodological problem as well. In clinical trials and long-term observational studies, patients with RA are generally assessed over the course of a few years, generating longitudinal data for all the attended visits. Finding a time-independent clustering would allow us to identify all potential clusters and their corresponding characteristics, including those patients whose residual burden is not captured by common disease activity measures. Considering the number of variables registered, a dimensionality reduction technique might be fruitful. The aim of such techniques is to decrease the number of variables whilst minimizing any loss of the original information [24]. Furthermore, these techniques are useful in uncovering latent, not directly observable, factors [14].

Once each patient's visit can be represented by a smaller but meaningful set of variables, the patients can be classified into a low number of clusters. Over time, different treatments might have different responses in the sample, and the patients' measurements fluctuate.

Therefore, the clusters' structure may remain unvaried while the patients may change cluster membership. Other possible results could exhibit that cluster solutions may be com-

pletely different at different time points, or, contrariwise, they could be reasonably constant over time, with just a few patients shifting the cluster membership.

The methodological problem entails the combination of the cluster analyses from different time points. To assess if the cluster solutions are clinically interesting, demographic, psychological, and health variables (including both the measurements and the treatment specification) can be used to investigate the association with the cluster memberships.

1.4 Goal of the study

Extensive literature exists about RA, with books and journals entirely dedicated to the subject. The focus is often on the risk factors, the symptoms, possible comorbidities, or the treatment strategies. All of these themes heavily rely on clinical and laboratory assessments. Recent studies started focusing more on patient-reported outcomes and considering the additional patients' needs. In the study described in Section 1.2, the patient-reported factor was shown to be predictive of sustained remission from as early as baseline [23].

There is, however, a lack of research on the disease activity of clusters of patients that have comparable measurements and responses to the treatment. This research gap limits our knowledge of a time-independent grouping of patients with similar characteristics. There is still a need for research on the clustering of individuals with similar RA disease activity. Such findings could help rheumatologists understand the possible courses of disease activity based on the patient's membership to a certain cluster.

This study aimed to identify a clustering of RA patients that is time-independent, within the time frame of the clinical trial considered. For this study we did not neglect the effect of time, nor did we expect clusters to remain unvaried with respect to the patients' memberships. Contrariwise, this study allowed the identification of all clusters that might arise when considering a similar study population [1]. An exploratory factor analysis similar to the one carried out by Pazmino et al. was implemented [14]. This uncovered the three factors: the patient assessment, the clinical assessment, and the laboratory assessment. By means of the three factors, the patients were clustered. We were interested in finding a clustering of patients with distinctive characteristics. Such classification could be helpful to rheumatologists in predicting sustained remission and quality of life in the medium term. The study consisted of two different but related analyses.

1. Cross-sectional analyses.

   This study employed data from the early stages of the patient's disease. We were interested in finding clusters among patients by employing only the observations from the beginning of their clinical trial. We used data both from Baseline (before the patients started their treatment) and from the firstly scheduled visit after the treatment had begun, at Week 8. The analyses were carried out separately obtaining two different cross-sectional analyses. We theorized that patients could be classified into a low number

[1] The terms "study population" and "sample" will be used interchangeably throughout the thesis, as they convey equivalent meanings within clinical terminology.

of groups with shared traits. To determine the clinical relevance of the cluster solutions, we examined socio-demographic and health variables, as well as treatment strategies.

2. Time-incorporated analysis.

   The main goal of the study was to build on the findings of the cross-sectional analyses and apply them to longitudinal data. We hypothesized that there exists a grouping of patients that is time-independent and describes the patient's outcomes at least throughout the first couple of years after the beginning of their treatment. The analysis incorporates different time points and is thus referred to as time-incorporated. The number of clusters found in the previous study can be used as a starting point to evaluate the sample's groups throughout the trial period.

For this purpose, the following research questions have been formulated, one for each part of the study.

1. What characteristics define the clustering of patients at the start of their treatment?

2. What combination of factors best describes a clustering of the patients that is time-independent, within the time frame of the clinical trial?

With this study, I wish to contribute by providing general guidelines that could be used by rheumatologists when determining the optimal T2T strategy for their patients. A definite cluster solution might help rheumatologists predict the course of a patient's disease activity based on their measurements and the different treatment strategies available.

Given the relatively small size of the study population considered, it is not feasible to split the patients into different samples to obtain formal validation. We call this a proof-of-concept study: once the methods have been proven to work and have been coded, they can be easily applied to other datasets. Once the results are validated, they could be useful for obtaining cut-offs for the factor scores, thus providing a "traffic light system" that aids rheumatologists in their decision-making.

1.5 Structure of the thesis

This thesis is intended to be accessible and informative for a diverse audience, including both statisticians and rheumatologists. Whilst some concepts may appear obvious to either one of the groups, we decided to include any explanation that might make the thesis more accessible to the other.

In Chapter 2 we introduce the motivating case study alongside the dataset that was used for the research. We present the study population's characteristics and the study design. The variables used in the analyses are also listed, with a short description for each variable. These are common measures collected routinely with RA patients, thus their explanations might be more informative to readers who are not familiar with rheumatology.

As the study consisted of two analyses, these are split into different chapters. Chapter 3 encloses the methods used for the cross-sectional analyses. The details regarding the multiple imputation procedure, exploratory factor analysis, and the clustering of the patients are included. Both the methods and their application to the dataset are explained. When a specification of a method did not apply to our dataset, this is not explained further, but relevant literature is included for the reader who is interested in additional information. The results of the exploratory factor analyses are also included in this chapter as they are necessary for the remaining parts of the methods' description. The interpretation of these results, on the other hand, is deferred to the following chapter.

The results of the first study can be found in Chapter 4. These precede the methods of the time-incorporated analysis, as the second analysis is built on the findings of the first. The product and interpretation of both the exploratory factor analysis and the cluster analyses are included. Furthermore, the clinical interpretation of the clusters is specified.

In Chapter 5, the methods from the time-incorporated analysis are explained. These are built on the methods from the previous study, thus a few elements are shortly repeated but without a detailed explanation. The differences in the application of the methods from the previous study are specified. The main focus of the chapter is on the methods that have not been illustrated yet, such as multiple outputation. Rather than a formal analysis, the second study was designed to be a proof-of-concept, primarily aiming to assess the effectiveness of the methods.

The results of the second study are enclosed in Chapter 6. The results and explanations for both the exploratory factor analysis and the cluster analysis are provided. While the outcomes of the study and the clinical interpretation of the results provide a clear image of the selected datasets, a formal validation on a new study population is needed. Once the results are validated, the methods could be applied to larger and more heterogeneous datasets.

In Chapter 7, a comprehensive discussion of the overall research is outlined. The results of the two studies are summarized, providing answers to the research questions, while relating the findings of the analyses to outcomes observed in earlier studies. We outline the impact and reasoning behind some of the statistical models, and present various alternatives. The strengths and limitations of the research are delineated. Lastly, possible future work is described.

Chapter 2

Study design

2.1 Motivating case study

Care in Early Rheumatoid Arthritis (CareRA) was a prospective 2-year investigator-initiated randomised pragmatic open-label superiority trial (EudraCT-number 2008-007225-39; ClinicalTrials.gov: NCT01172639) [25]. The trial was conducted in 13 Flemish rheumatology centers (two academic centers, seven general hospitals, and four private practices) in Belgium. The trial was carried out between 2009 and 2015.

2.2 Patients

400 patients with recently diagnosed RA (disease duration up to one year) were recruited. Of these, 379 were included in the analysis, with the remaining 21 patients excluded due to screening or randomisation errors [26]. Patients were excluded from the trial if they had previous treatment with certain drugs that would interfere with the treatment in the clinical trial, and if they had psoriatic comorbid arthritis. Other comorbidities that would create an unacceptable risk for participation were also a cause of exclusion. Lastly, patients who were pregnant, breastfeeding or diagnosed with substance use disorder were also excluded. Women constituted 69% of the patients, and the average age at the end of the trial was 53.9 years (SD: 13.0, Range: [21-82]).

2.3 Procedure

Patients were stratified into a high-risk or low-risk group. The allocation was implemented according to standard prognostic markers such as the presence of erosions, and the baseline DAS28 [27]. The allocation resulted in 289 high-risk and 90 low-risk patients.

The 289 patients stratified into the high-risk group at screening were randomised into three treatment strategies: COBRA-Classic, COBRA-Avant-Guarde, and COBRA-Slim[1]. The 90 patients stratified into the low-risk group were randomised into two treatment strategies: Tight Step-Up, and COBRA-Slim. The latter was equivalent to the third treatment strategy from the high-risk group. Patients were assessed at screening and baseline and then followed up at Week 8, 16, 28, 40, 52, 65, 78, 91, and 104. Demographics were registered at screening,

[1]The names of the treatment strategies are included for completeness, but their efficacy and effectiveness are beyond the scope of this study. Further information can be found in the paper by Verschueren et al. [26]

while clinical and laboratory parameters along with patient variables were collected at every visit [14], [25].

Over the course of the trial, various measures were evaluated. In particular, following the study by Pazmino et al, nine variables were chosen for the analyses. Overall, only 14% of the values are missing from the dataset with all the patients measured at every scheduled visit. In general, data is missing on a visit level: in most cases, participants miss all the measures from a visit. When only a few measures are missing from a visit, these are mostly the ones related to the patient's perception of the disease. The patient might have considered them less important in the diagnosis, or could have been fatigued by the other tests.

Another possible source of missingness could be the death of some participants. Only two participants died during their follow-up period. Since RA leads to a life expectancy that is 3 to 10 years lower than the general population, there could be numerous reasons for such a low rate. Firstly, the patients had all been diagnosed with RA in the year prior to the beginning of their treatment. An early diagnosis helps rheumatologists take action before the disease activity causes functional impairment or an excessive burden on the patient. Furthermore, the study had a duration of only two years, meaning that at the end of the trial, the patients had been diagnosed with RA for a maximum of three years. The study population was also relatively young, with a mean age of 53.9 at the end of the trial, against the average age of onset set at 58 years. Lastly, the trial employed state-of-the-art medication, with a possible treatment adaptation for those who did not respond adequately to the initial treatment scheme [14]. These reasons might explain the very low death rate during the trial period.

2.4 Variables

Many different variables associated with the disease activity were collected at every visit. In the study by Pazmino et al., nine variables were employed in the exploratory factor analysis. These are:

- C-reactive protein (CRP)
  The CRP is an acute-phase[2] reactant present in the blood plasma. Its levels increase in response to inflammation and are a marker of systemic inflammation in RA. CRP plays a substantial part in the defence mechanism against infections [28], [29]. In RA, CRP has been associated with comorbidities such as pulmonary diseases, cardiovascular diseases, and depression. In general, greater levels of CRP are associated with higher RA disease activity. In healthy subjects, CRP concentration is usually lower than 10 mg/L. Levels above this threshold are usually considered elevated. An inflammatory event can bring the levels to soar in a few hours, reaching a concentration between 20 and 500 mg/L [30]. In general CRP levels are higher in RA patients, with a concentration higher than 20 mg/L often reported at baseline [29], [31]. In our study population, 299 of the 379 patients (79%) presented a CRP level of less than 10 mg/L at baseline. The

[2]Stage of the disease characterized by active inflammation, often marked by symptoms such as joint pain, swelling, stiffness, and fatigue

reason for such a difference could be the selection criteria which only included patients with early RA. Nonetheless, observational studies reported that many RA patients still exhibit normal levels of CRP [32], [33]. Therefore, the measure only reflects parts of the disease activity and needs to be assessed in a wider framework [29].

- Erythrocyte sedimentation rate (ESR)
  The ESR is a simple laboratory test, that proved to be useful in the diagnosis of inflammatory activity [34]. The test requires 2 ml of blood to be poured into a vertical tube called the Westergren tube. Due to the effect of gravity, the red blood cells are deposited at the bottom of the pipe, separating themselves from the plasma. The ESR measures the height in mm of the red cell sediment exactly one hour after the blood has been poured. Following an inflammation, the red blood cells clump together, making them heavier. This results in faster separation and settlement which causes higher ESR levels. The reference range for the measure is between 0 and 15 mm/hr for men, and between 0 and 20 mm/hr for women [35]. At baseline, only 21% of the study population displayed ESR levels within the normal reference range.

- Swollen joint count (SJC)
  The assessment of the joints is an essential element in the evaluation of RA disease activity. Thus, counting the number of joints affected by the disease is a reliable metric. When measuring the SJC, rheumatologists inspect the soft tissue swelling and effusion in the joints [31]. Studies used to advise the inspection of 68 joints [36]. After the formulation of the DAS28 as a measure of the disease activity, a simplified 28-joint count proved to provide comparable data to the one obtained when employing the more comprehensive measure. Furthermore, the simplified count requires 60% less time to execute when compared to the higher joint count [18]. The SJC measures the number of swollen joints among the 28 ones considered in the DAS28. For this reason, it can also be named SJC28.

- Tender joint count (TJC)
  Similarly to the SJC, the TJC aims to assess the number of joints affected by tenderness on pressure or motion. This also considers only 28 joints and can be named TJC28 [31].

- Pain
  Pain is one of the notable symptoms and residual burden in RA. When dealing with a perception or a feeling it is hard to assess this through an objective evaluation. The variability of pain is best assessed through a continuous scale. A discrete scale, with increasing numerical values or categories, would not mirror the variability of the feeling. For this reason, usually pain is measured on a visual analog scale (VAS) [31]. This is a 100 mm line that depicts the continuous range between the two extremes (no pain and worst pain). The patient is asked to self-report their perception of pain by marking a point on the line that is proportionate to the feeling. The measurement is then determined by calculating the distance in millimeters from the left endpoint to

the patient's mark [37]. Generally, the perception in the week preceding each visit is evaluated.

- Fatigue
  The perception of fatigue is evaluated equivalently to the perception of pain. A 100 mm VAS is employed in order to measure the feeling over the week preceding the visit.

- Patient global health assessment (PaGH)
  PaGH is a common patient-derived element in the evaluation of the disease activity score. Patients are asked to answer the question "Assuming all the ways your life is affected by your rheumatism, how did you feel on average over the past week?" [14]. Once again, a 100 mm VAS is employed. The PaGH is an essential component of compound disease activity scores, as it voices the patient's global health perception. However, the broad formulation is not unambiguous. A study showed that patients generally do not know the functionality of this assessment. Whilst some believe it is just a general question about their health, it can actually impact the treatment strategy [38].

- Physician global health assessment (PhGH)
  While PaGH is a more subjective measure derived from the patient's perception of their own global health, the PhGH is observer-derived. The two are in general highly correlated, with the former frequently receiving higher ratings [31]. The physician was asked to evaluate their perception of the patient's global health on a 100 mm VAS.

- Health Assessment Questionnaire (HAQ)
  This questionnaire is commonly employed to assess the ability to perform daily life activities, and is used in most RA trials [39]. The assessment consists of 20 questions aimed at evaluating the upper and lower extremities. The questions are arranged into 8 categories: rising, eating, dressing, walking, hygiene, reach, grip, and usual activities. Patients answer on a Likert scale. For each question, the patient indicates the grade of difficulty they encounter in a particular activity. The scale ranges from no difficulty, some difficulty, much difficulty, and inability. Each answer is converted into a number on a scale ranging from 0 to 3. The final HAQ score is computed by calculating the mean of the highest score in each of the 8 categories. Since the answers range from 0 to 3, so does the HAQ score, with higher scores expressing a higher degree of disability. Levels below 0.3 are considered to be close to normal [31].

While the previous variables were collected at every visit, more variables were collected at screening only. Some of these were employed for the stratification of the patients into the two risk groups. In Section 3.1 we will explain how multiple imputation was carried out in order to obtain a complete dataset. Among the set of variables collected at screening, the following six variables were used to increase the accuracy of the imputations:

- Age (in years)

- Sex (female / male)
- Center of recruitment
  This refers to the 13 Flemish rheumatology centers where the trial was conducted.
- Rheumatoid factor (RF)
  The rheumatoid factor is a type of autoantibody found in the serum, the liquid component of plasma. Only 70 % of RA patients exhibit the presence of RF. Furthermore, 15% of the general population has RF without actually being diagnosed with RA [40].
- Anti-cyclic citrullinated peptide antibody (ACPA)
  ACPA is another antibody found in the serum, used for the diagnosis of RA. It can be found in only 3% of the general population. The occurrence of either or both RF and ACPA is associated with a higher disease severity [40]. While ACPAs seem to offer greater specificity in diagnosing RA, RF remains a key indicator for the diagnosis, advancement, and severity of RA [41].
- Erosions

  Bone erosions are a crucial factor in assessing RA and serve as a predictive indicator for more severe disease activity, leading to increased disability and elevated mortality rates. Therefore, currently, identifying and measuring bone erosion plays a significant role in diagnosing the disease and evaluating the effectiveness of drug therapy in RA patients [42].

Chapter 3

Methods: Cross-sectional study

The two cross-sectional analyses employed data exclusively from Baseline and Week 8 separately. As a result of the treatment effect, many patients had lower measurements during the visits succeeding Baseline. Particularly, Pazmino et al. noticed that both the clinical, laboratory, and patient-reported factors showed significant improvement within the initial 8 weeks, and remained overall stable after that time point [23]. From a clinical perspective, we could not combine the data from before and after the treatment had begun. Hence, two separate cross-sectional analyses were carried out for the two initial visits. For both visits, a cluster solution was found. The visits succeeding week 8 were analysed in the subsequent study.

At the beginning of the trial, the percentage of missing data was at its lowest. At Baseline, all patients were evaluated. Missingness arose only on a variable level, with a few participants not responding to some of the questions about their perception of the disease (such as pain or fatigue), and some participants not having their ESR measured. In general, of all the Baseline values, less than 0.5% were missing. At Week 8, 12 participants did not attend the scheduled visit. Again, incompleteness was still very low, with less than 3.5% of the values missing from the dataset. Upon consultation with the clinical team that provided the data, missing data were assumed to be missing at random (MAR)[1], meaning that the likelihood of a missed visit was not related to the patient's disease activity [14].

All analyses were performed with R (version 4.1.1).

3.1 Multiple imputation

In clinical studies like CareRA, data collection might be affected by missing data due to item non-response, when only some variables are missing, or missed visits. In these contexts, a common approach for handling missingness is through multiple imputation (MI). A benefit of such a technique is that it allows to perform standard analysis on a complete dataset, without resulting in loss of information from the deletion of incomplete observations [45]. Furthermore, differently from other imputation techniques, MI accounts for the uncertainty of the imputed values for the missing observations.

The underlying idea beneath the MI procedure is the replacement of each missing value with a series of m possible values. The imputed data are a Bayesian draw from the distribution

[1] This is one of the mechanisms behind missing data first introduced by Rubin to explain the relation between the missingness indicator and the data [43]. Additional information can be found in the paper by Schafer and Graham [44].

of the missing values conditional on the observed values. The procedure generates m complete datasets, which can be employed for further analyses. MI is commonly used under the MAR assumption, as in our case [46].

A frequently used approach for implementing the process is through multiple imputation by chained equations (MICE) [47]. The missing values of a variable X_1 are imputed from a regression of the observed values of X_1 on (X_2, X_3, etc.). Likewise, the missing values of X_2 are imputed from a regression of X_2 on (X_1, X_3, etc.), and so forth [45].

A drawback of MICE is the specification of conditional models for all the variables that are not complete. Identifying the possibly non-linear relations and interactions among the variables could be burdensome. A solution is to use classification and regression trees (CART) as conditional models [48].

CART models aim to estimate the likelihood of a single variable based on several predictor variables. The CART algorithm divides the predictor space into segments, ensuring that each segment contains relatively similar outcomes. These segments are identified through a sequence of binary splits of the predictor variables [45].

A CART model is beneficial for our specific type of data as well. Since the imputations originate from the observed values, any restrictions would be enforced systematically. For example, variables measuring the joint counts (SJC28 and TJC28) would automatically have integers ranging between 0 and 28 as imputed values. Likewise, the imputed HAQ scores would only range between 0 and 3 as there was no observed score outside this range.

Consider the subgroup of variables used for the cross-sectional analysis. The nine variables employed for factor analysis were measured both at Baseline and at Week 8. Whilst many other measures were collected at screening, following the procedure carried out by Pazmino et al. six additional variables were employed to achieve more reliable imputations. Thus we obtain a matrix with 379 patients and 24 variables (the nine variables measured both at Baseline and at Week 8, plus the six variables used for the imputation model). Our 379×24 matrix X could be constructed so that the set of j columns X_J with missing values are to the left of the set of columns X_K that are entirely observed. The matrix is denoted as $X = (X_J, X_K)$. We also assume that, when scanning X_J from left to right, the percentage of missing data in each column is nondecreasing [45].
Implementing MICE with CART as the conditional model for imputation was accomplished following a 5-step algorithm [45].

1. Define a matrix Z and set it equal to the complete part of the original dataset X_K.

2. Fill in the initial values for the missing elements: for $i = 1, .. j$, the missing values in each column X_i are imputed using a CART on Z. X_i is then appended to Z before incrementing i.

3. For $i = 1, .. j$, replace the formerly missing element of X_i by using a CART on Z_{-i}, where Z_{-i} is the matrix Z with the i^{th} column eliminated.

4. Repeat step 3 l times.

5. Repeat steps 1-4 m times, obtaining m imputed datasets.

Since setting $l = 10$ generally produces satisfactory results, this value was chosen [45]. Considering the low percentage of missingness, we set m equal to 2, obtaining 2 imputed and complete datasets. The two datasets were inspected to assess whether their dissimilarity established the need for a higher number of imputations. Descriptive statistics for each variable across the two imputations were compared and no notable difference was reported. The minimum leaf size was set to 5. Moreover, if the deviance of its values is smaller than 0.0001, a leaf was not split further [45].

3.2 Dimensionality reduction

For each of the two visits considered in this study, two complete datasets were obtained. Each dataset comprised the nine variables defined in Section 2.4. Although this was already a subset of the numerous variables collected at each visit, in order to perform cluster analysis on the observations, a dimensionality reduction proved to be useful. Moreover, some of the variables' descriptions were logically related. For example, with a high value on the TJC28, one expected a similar value on the SJC28. Equivalently, the values measured on the VAS (such as pain, fatigue, or the patient global health assessment) were expected to be reasonably correlated. A table containing the Pearson correlation of the nine variables in our dataset can be found in the paper by Pazmino et al. (p. 177, Table 2) [14].

A dimensionality reduction technique aims at diminishing the number of variables while retaining as much of the information as possible. The goal is to create n new variables from the set of p original ones, with limited loss of information. Both principal component analysis (PCA) and exploratory factor analysis (EFA) are common techniques used for this purpose.

PCA aims to explain the maximum possible variance within the data with as few composite variables as possible. The components uncovered through PCA are ordered decreasingly with respect to their variance explained and are always uncorrelated to each other. Thus, a component z_i has the largest variance among all the possible linear combinations of the variables x_1, x_2,..., x_p whilst remaining uncorrelated to z_1, z_2,..., z_{i-1}. This guarantees each component to account for a different source of variation. Because of this characteristic, the interpretation of the components might not always be straightforward [24].

The principal components are computed by identifying the eigenvectors and eigenvalues of a matrix. In general, the covariance or the correlation matrices of the original variables are employed. In our case, the latter was preferred, since when adopting the covariance matrix we assume all the variables to be measured on the same scale and have a comparable range of variation. When employing the correlation matrix no such assumption is required.

Like PCA, EFA aims to reduce the dimensionality of the dataset. However, the idea behind EFA is that there exist latent unobserved variables underlying the observed ones. EFA aims to uncover these common factors, and the focus is on the interpretation of the underlying factors, rather than the explanation of the variation of the data. For this reason, especially in medical literature, EFA is often preferred as it allows for a clearer description of the factors. Lastly, dissimilar to PCA, the factors do not necessarily have to be uncorrelated [24].

The factors are obtained through the extraction of the correlation matrix eigenvalues.

Depending on the data type, a certain extraction technique might be preferred. The most common ones are the generalized least-square method, the maximum-likelihood method, and the principal component extraction [49].

3.2.1 Exploratory factor analysis with principal component extraction

The difference between PCA and EFA with principal component extraction is in the step subsequent to finding the initial components. PCA aims to maximize the variance of the first components, and halts once an appropriate number has been extracted. With EFA, the assumptions of decreasing variance and independence of the factors do not hold anymore. Hence, any solution found through EFA is not unique. After having obtained the matrix of factor loadings Λ, this can be multiplied by an orthogonal matrix T, creating a new solution that is also optimal. This multiplication is identical to an orthogonal rotation of the axes. A generalization of this rotation allows oblique rotation, creating factors that are non-orthogonal [24].

In exploratory factor analysis, a solution produces the factor loadings. These are the "correlation of the original variable with a factor" [50]. After having found a solution, this can always be rotated. The purpose of the rotation is the simplification of the structure. Hence, among all possible solutions, the preferred one is the one that is the clearest to interpret[2]. To accomplish this, it is necessary to rotate the factors so that each variable would load onto as few factors as feasible.

Hence, the goal of the rotation is to have variables correlate strongly with one factor and weakly with all the other factors. When the correlation between a variable and a factor is below 0.3, it is considered negligible and can be disregarded [52, pag. 210]. Optimally, we aim to obtain a matrix where each variable has a correlation above 0.3 with only one factor. Whenever a variable correlates strongly with more than one factor, the factor with the highest correlation is called the primary loading, whilst the remaining are named cross-loadings [53].

The matrix rotation can be orthogonal or oblique. In orthogonal rotation, the extracted factors are not correlated with each other, whereas in the case of oblique rotation, the assumption of independence does not hold. Determining the appropriate rotation type is dependent on whether there exists a reason to assume a relationship or independence between the factors, as well as how the variables are grouped together on the factors before the rotation [54]. The most common types of rotation are varimax, equamax, quartimax, promax, and direct oblimin. The former three are examples of orthogonal rotation, whilst the latter two are instances of oblique rotation. In our case, there was no reason to assume independence of the factors, so an oblique rotation was chosen. Specifically, promax was selected as it delivered a matrix with the smallest number of cross-loadings.

The promax rotation transforms an orthogonal rotation into an oblique solution. The orthogonal solution is employed as a foundation for generating an optimal oblique solution. Generally, either varimax or quartimax are used as the initial solution [52, p. 190]. To obtain

[2]Whilst this can be a subjective interpretation, there are general guidelines that define the criteria for a simple structure. These can be found on page 335 of the book by Thurstone [51]

a solution better than that obtained by the orthogonal rotation, it is necessary to reduce the lower loadings, while simultaneously maintaining relatively high primary loadings. This enhanced solution becomes feasible by allowing the factors to be oblique. From a mathematical perspective, raising all factor loadings to a higher power results in lower values[3].

As the exponent increases, the loadings naturally diminish, consequently resulting in increased correlations among the factors. By adjusting the power, it is possible to control both the level of obliquity and the simplicity of the structure. The optimal power is the one that produces the most straightforward structure with minimal correlation among factors. Typically, a satisfactory solution is obtained by applying a power of four to the loadings, although sometimes a power of two is employed when it appears to yield a better structure [52, pp. 190-191].

3.2.2 Number of factors

Another important distinction between PCA and EFA is in the choice of the number of components to extract. In PCA, when uncovering *(m+1)* components, the variance explained by the new component *(m+1)* is simply added to the variance explained by the m previous ones. In EFA, on the other hand, the process of rotation reassigns the overall variance among factors. In this case, transitioning from m to *(m+1)* factors has the potential to entirely alter the characteristics of each factor [24]. Consequently, the choice of m often holds greater significance than the choice of the rotation method in determining the nature of the factors.

Until now, we assumed to perform EFA with a known number of factors, whilst this is often not the case. This study was partly a replication of the study performed by Pazmino et al., where a three-factor model proved to be optimal, thus we decided to extract 3 factors [14].

3.2.3 Exploratory factor analysis on multiply imputed datasets

A problem common to both PCA and EFA is the application of the techniques to multiply imputed datasets. Both methods employ eigenvalue decomposition of either the covariance or the correlation matrix [4]. Consider, for example, a dataset X_{obs} with missing values, and the datasets X_1, X_2,..., X_M as the complete imputed datasets. When performing PCA on the correlation matrix of X_1, the eigenvector associated with the highest eigenvalue describes the structure linked to the first latent factor. When performing PCA on the correlation matrix of X_2, there is no assurance that the eigenvector associated with the highest eigenvalue of

[3]For instance, if the initial loadings are 0.9 and 0.3, the latter is one-third the size of the former. However, when squared, the loading for the second variable (0.09) is one-ninth the size of the squared loading for the first variable (0.81). Consequently, the absolute difference also grows. In order to further amplify the relative difference between the high and small loadings, the factor loadings can be raised to a power greater than 2. As the power increases, the lower loadings approach zero [52, p. 190].

[4]As explained in Section 3.4, in our case, the factors are computed to cluster the patients. In general, performing PCA or EFA on the average correlation matrix would make it infeasible to apply combination rules for precision-related objectives. As it does not apply to our case, this specification is not analysed any further. Additional information can be found in the paper by Nassiri et al. [55]

$\text{Corr}(X_2)$ is directly comparable to the one obtained from $\text{Corr}(X_1)$ [55]. More precisely, computing the mean of the eigenvectors based on the order, or the eigenvalues derived from the correlation matrix of each imputed dataset, is likely to produce meaningless results [5] [55].

A solution to this problem would be to determine the correlation matrix from the imputed dataset employing Rubin's rules for multiple imputation. Consider again the M imputed dataset X_1, X_2,..., X_M, with correlation matrices $\widehat{Corr(X_1)}$, $\widehat{Corr(X_2)}$,..., $\widehat{Corr(X_M)}$. A multiple imputation estimate of the population correlation matrix can be obtained by averaging the correlation matrices of the imputed dataset [56]. Thus:

$$\widetilde{Corr(X)} = \frac{1}{M}\sum_{i=1}^{M} \widehat{Corr(X_i)}$$

After having obtained $\widetilde{Corr(X)}$, either PCA or EFA can be performed on it. This would eliminate the issues of factor ordering and the need to determine the number of factors across multiple imputation.

3.3 Factors scores

The three factors extracted were equivalent to the ones described by Pazmino et al. in their study where the factors were obtained using all ten visits [14]. These were the patient-reported (PRF), clinical (CF), and laboratory factors (LF). As the factor loadings are necessary for the subsequent explanation of the factor scores, these are already reported in Table 3.1. This includes two equivalent sub-tables, each containing the results of the EFA on either the observations from Baseline or Week 8. The factor loadings exceeding the 0.3 threshold are displayed in bold. While in the analysis employing data from Week 8 there is no cross-loading, at Baseline PhGH loads both onto the patient-reported factor and the clinical factor. Further discussion on the factor loadings' description and interpretation is deferred to Chapter 4.

As discussed in Section 3.1, two complete datasets were obtained. In order to obtain the factor scores for each patient and perform a cluster analysis on the results, the two datasets were averaged, obtaining a single dataset comprised of the nine variables measured at two time points for each patient.

The variables were scaled so that each variable would have a range between 0 and 100. Naturally, the variables that were measured on the VAS (pain, fatigue, PaGH, PhGH) were unaffected as the range is equivalent. The HAQ and the two variables measuring the joint count (SJC28 and TJC28) ranged between [0-3] and [0-28] respectively before the transformation.

[5] Whilst this does not apply to our case, an additional challenge would be the choice of the number of factors. It is essential to establish a consistent number of factors across multiple imputed datasets. Neglecting to do so might result in a different number of factors uncovered for each dataset, which would hinder the combination of the results. However, there is no assurance that the methods used to determine the optimal number of factors would yield the same result for each imputed dataset.

Baseline				Week 8			
Variables	Factor 1: Patient	Factor 2: Clinical	Factor 3: Laboratory	Variables	Factor 1: Patient	Factor 2: Clinical	Factor 3: Laboratory
Pain	**0.96**	-0.06	-0.02	Pain	**0.87**	0.04	0.02
PaGH	**0.93**	-0.04	0.05	PaGH	**0.90**	0.04	-0.08
Fatigue	**0.89**	-0.13	-0.07	Fatigue	**0.94**	-0.26	-0.02
HAQ	**0.59**	0.19	0.11	HAQ	**0.62**	0.21	-0.04
SJC28	-0.14	**0.97**	0.04	SJC28	-0.20	**0.98**	0.06
TJC28	-0.08	**0.96**	-0.04	TJC28	0.02	**0.89**	-0.08
PhGH	**0.34**	**0.64**	-0.03	PhGH	0.22	**0.68**	0.03
CRP	-0.04	0.01	**0.89**	CRP	0.02	0.03	**0.84**
ESR	0.03	-0.02	**0.89**	ESR	0.01	-0.01	**0.85**

Table 3.1: Factor loadings of the three factors (correlation between the observed variable and the latent factor) for both Baseline and Week 8, separately. Displayed in bold are the loadings exceeding the 0.3 threshold. The factors are ordered by the percentage of variance explained.

A different approach was taken for CRP and ESR. Considering these are biomarkers found in the blood, there is no natural maximum value that can be used for scaling the variables. Furthermore, both variables were very skewed, with the majority of the patients exhibiting relatively low values. For example, we have already illustrated that 299 of the 379 patients presented a CRP level of less than 10 mg/L at Baseline. On the other hand, the maximum value for CRP registered at Baseline was 186.1 mg/L. Thus, if we were to employ such a high value for rescaling the variable, the vast majority of the observations would have had a very low CRP level. This might have hindered the clustering solution which would have presumably clustered these patients together, disregarding the relative differences among patients with lower values.

This brought us to consider an alternative solution for CRP and ESR. Both variables were examined through a series of visualization techniques. Specifically, a boxplot of the CRP measurements at Baseline revealed that the upper quartile had a value of 8.4 [6]. This means that 75% of the observations had a value of at most 8.4. Another useful feature of a boxplot is the upper whisker[7]. When analysing normally distributed data, any observation above this level is usually considered to be an outlier. At Baseline, the upper whisker for the

[6]The boxplot can be found in Appendix B at page 63.

[7]This denotes the upper quartile times 1.5 times the interquartile range (i.e. the range of the data that goes from the lower quartile up to the upper quartile).

CRP measurement had a value of 20, with 44 patients exhibiting a larger measurement.

Whilst the observations were not normally distributed, and a relatively large number of patients exceeded that value, 20 was still chosen as an upper threshold for the scaling. Thus, any observation above that value was automatically assigned a value of 100, and the rest was scaled accordingly. The reason for this choice is two-fold. Firstly, as explained in Section 2.4, a CRP level above 20 mg/L might denote an inflammatory event, which brings levels to soar for a few hours. Secondly, the distribution of CRP at every visit, including the ones after Week 8, was inspected. Slightly more than 3% of the observations exceeded the threshold of 20. If a higher threshold were to be chosen, the relative differences between observations with lower values would decrease further. Such a transformation would result in a cluster solution that focuses more on the few outliers, rather than the differences between the majority of the patients with low CRP values.

A similar analysis was carried out for the ESR. At Baseline, the upper whisker of the corresponding boxplot had a value of 88, with just 11 patients displaying a level greater than the threshold. These were automatically rescaled to 100, while the rest of the patients were rescaled from a range of [0-88] to a range of [0-100].

To compute the factor scores for each patient, the scaled variables were multiplied by their factor loadings and summed up so that each factor would comprise the variables that loaded onto it[8]. Employing the results from Table 3.1 at Baseline the three factor scores were computed in the following way:

$$PRF = (0.96 * Pain) + (0.93 * PaGH) + (0.89 * Fatigue) + (0.59 * HAQ)$$
$$CF = (0.97 * SJC28) + (0.96 * TJC28) + (0.64 * PhGH)$$
$$LF = (0.89 * CRP) + (0.89 * ESR)$$

Equivalently, in Week 8, the factor scores were computed as follows:

$$PRF = (0.87 * Pain) + (0.90 * PaGH) + (0.94 * Fatigue) + (0.62 * HAQ)$$
$$CF = (0.98 * SJC28) + (0.89 * TJC28) + (0.68 * PhGH)$$
$$LF = (0.84 * CRP) + (0.85 * ESR)$$

Due to the different number of variables for each factor, each factor score was standardized to a [0–1] scale (with higher values indicating a greater health impact). Consequently, for both visits analysed, each patient had three factor scores ranging between 0 and 1 [23].

[8]The decision to exclude PhGH from the computation of the patient-reported factor, despite the loading exceeding the 0.3 threshold, ensued the subjective preference to solely include the primary loadings. Furthermore, this part of the analysis was partly a replication of the study by Pazmino et al. [14]. In their study, the researchers carried out a longitudinal analysis of the data from Baseline until Week 104. The results displayed no cross-loadings. Lastly, considering the loading of PhGH onto the patient-reported factor just exceeded the threshold of 0.3, it was not taken into consideration.

3.4 Clustering

We hypothesized that patients might be categorized into a small number of groups characterized by common traits. The goal of cluster analysis was to group the patients into subsets or clusters. The aim of this process was to ensure internal coherence within each cluster while maintaining clear distinctions between clusters. Therefore, the patients within a cluster would have exhibited high similarity, while patients in different clusters would have been as dissimilar as possible [57].

At the core of cluster analysis lies the concept of the level of similarity (or dissimilarity) among the individual patients. A clustering approach seeks to group the patients based on a provided definition of similarity. The majority of clustering algorithms require a dissimilarity matrix as input [58]. In our case, a dissimilarity matrix was built upon the distances between measurements x_{ij}, with $i = 1, 2, ..., 379$ denoting each patient and j referring to one of the three aforementioned factors. One of the most common options to measure the dissimilarity between two objects is by computing their squared distance. Thus the distance between two patients was defined as:

$$d(x_{\mathrm{ij}}, x_{\mathrm{lj}}) = \sum_{j=1}^{3} (x_{\mathrm{ij}} - x_{\mathrm{lj}})^2$$

3.4.1 K-means

The K-means algorithm is one of the most common clustering methods. It is designed for situations where all variables are quantitative, as in our case, and uses the squared Euclidean distance as the measure of dissimilarity [59]. The following algorithm partitions the data points into k clusters:

1. Randomly position k points within the space corresponding to the objects being clustered. These points serve as the initial centroids for the groups.
2. Assign each object to the group with the nearest centroid.
3. After all objects have been allocated, update the positions of the k centroids by computing the average position of each group.
4. Iterate through steps 2 and 3 until the centroids stabilize.

3.4.2 Number of clusters

As evinced by the algorithm itself, algorithms like k-means clustering require the user to define the number of clusters in advance. Identifying the optimal number of clusters in a dataset becomes a key problem. Unfortunately, there is no universal solution to this problem. The optimal number is somewhat subjective and it is influenced by the similarity measurement method and partitioning parameters employed [60]. Various techniques exist to determine an optimal number of clusters. Specifically, we employed:

- Average silhouette method
  This method evaluates the effectiveness of a clustering by assessing how well each object is positioned within its cluster. For each object i, the silhouette is computed through the following steps [60]:

  1. Compute the average dissimilarity, denoted as a_{i}, for each object i. This is achieved by measuring the average squared distance between i and all other objects within the same cluster.
  2. For all other clusters, denoted as C, to which i does not belong, calculate the average dissimilarity, denoted as $d(i, C)$, between i and all observations within C. The smallest value among these $d(i, C)$ calculations is defined as $b_{\mathrm{i}} = min_{\mathrm{C}} d(i, C)$, where b_{i} represents the dissimilarity between i and the nearest cluster to which it does not belong.
  3. The silhouette width of an observation i is determined by the formula

  $$S_{\mathrm{i}} = \frac{(b_{\mathrm{i}} - a_{\mathrm{i}})}{max(a_{\mathrm{i}}, b_{\mathrm{i}})}$$

  The average silhouette method calculates the mean silhouette of observations across various k values. The optimal cluster number maximizes the average silhouette width across a defined range of values [60].

- Gap statistic method
  This method computes the total within-cluster variation for various values of k and compares them with the values under the null reference distribution of the data (i.e. random uniform distribution of points) [60]. The following algorithm computes the gap statistics [61]:

  1. For each possible value of k, cluster the observed data and compute the total sum of squares W_{k} around the cluster means.
  2. Create B sets of reference data with a random uniform distribution. For each possible value of k, cluster each of these datasets and compute the total sum of squares $W_{\mathrm{k,b}}$ around the cluster means.
  3. For each possible value of k, calculate the estimated gap statistic $Gap(k)$, and its standard deviation.

  $$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} log(W_{\mathrm{k,b}}) - log(W_{\mathrm{k}})$$

  This is a measure of the deviation under the null hypothesis of the observed W_{k} from the expected $W_{\mathrm{k,b}}$.
  4. The optimal value of k is the smallest one such that $Gap(k)$ is within one standard deviation of $Gap(k+1)$. Thus:

  $$Gap(k) \geqslant Gap(k+1) - s_{\mathrm{k+1}}$$

The optimal clustering is the one that maximizes the gap statistic. Therefore, the partition is the furthest from the random distribution of the data [60].

3.4.3 Hierarchical clustering

Alternative algorithms, other than K-means, are available to cluster data in subgroups. Another widely used method is the hierarchical clustering algorithm. The agglomerative approach of this method generates hierarchical structures where, at every level, clusters result from merging clusters from the preceding lower level. At the lowest level, each cluster consists of a singular observation. At the highest level, there exists a single cluster including every observation. The agglomerative hierarchical clustering starts from the lowest level. At each step, it iteratively combines a pair of clusters into a singular cluster. This results in a grouping at the next level with one fewer cluster. The pair selected for merging comprises the two clusters with the lowest dissimilarity [58].

Every level in the hierarchy displays a distinct grouping of the data into separate clusters. It is at the user's discretion to determine whether a specific level, if any, truly illustrates a "natural" clustering, one where observations within each group are notably more similar to each other than to those in different groups at that level. The previously mentioned Gap statistic method can be employed for this evaluation [58].

3.4.4 Hierarchical K-means clustering

One drawback of the K-means algorithm is that it is sensitive to the initial selection of centroids. The final clustering solution might differ based on the location of the clusters' centroids during the first step of the algorithm [60]. A possible solution to minimize the bias generated from a random allocation of the centroid is to combine both hierarchical and K-means algorithms into one:

1. Identify an optimal number of clusters k.
2. Perform hierarchical clustering on the dataset and partition the structure into k clusters.
3. Compute the centroids of each cluster by calculating the average position.
4. Implement the K-means algorithm employing the set of cluster centroids determined in step 3 as the initial cluster centers.

The cluster analysis was performed on data from both visits, separately. At first, the number of clusters for each analysis was chosen. As both the clinical, laboratory, and patient-reported factors showed significant improvement within the initial 8 weeks, there was no expectation that the number of clusters had to be constant for the two analyses [23]. Therefore, the methods to identify an optimal number of clusters were applied to both visits independently. Once k was found, hierarchical K-means clustering was implemented in the two separate datasets. Once again, there was no expectation that the clusters found had to be constant at the two time points.

Chapter 4

Results and clinical interpretation: Cross-sectional study

All 379 patients included in CareRA was considered in the analyses. The average age at the end of the trial was 53.9 years (SD: 13.0) with patients ranging between ages 21 and 82. 69% were women, 77% either displayed positive rheumatoid factor or positive ACPA (both are variables associated with RA), and 19% presented other types of comorbidities such as cardiovascular diseases.

4.1 Factors

Two exploratory factor analyses were carried out, for Baseline and Week 8 separately. In Table 3.1, the first factor extracted explained the most variance for both the analyses, explaining 34% and 33% of the variance at Baseline and at Week 8 respectively. It included Pain, PaGH, Fatigue, and the HAQ. All these variables were patient-reported outcomes. Thus, following the nomenclature from Pazmino et al., it was denoted as the patient-reported (or patient) factor (PRF) [14]. The second factor extracted contained all the variables evaluated by the clinician: SJC28, TJC28, and the PhGH. For this reason, it was referred to as the clinical factor (CF). Finally, the third factor contained the two laboratory measures: CRP and ESR, and was thus designated as Laboratory factor (LF). The first latent factor comprises the perception of the disease burden by the patient. Particularly, it describes the means by which the disease activity influences the patient's perceived health and regular functioning. The latter two factors, on the other hand, describe the disease activity in terms of the biological inflammation in the peripheral joints [14].

At Baseline, the three-factor model explained 78% of the variance of the disease activity measured by the nine variables included. At Week 8, the variance explained was 75%. The factors in Table 3.1 were ordered by the percentage of variance explained, while the variables were ordered by their correlation with the primary loading at Baseline.

While it is not possible to compare directly the homonymous factors from the two analyses, the coefficient of congruence can be used to relate factors through their loadings [52, p. 285]. This coefficient, commonly known as Tucker's congruence coefficient, can be used to evaluate the similarity between multiple factor interpretations. Modern literature suggests that coefficients with a value between 0.85 and 0.94 exhibit a fair similarity. Values larger than 0.95 indicate the two factors to be considered equal [62]. Table 4.1 displays the coefficients of congruence between the three factors uncovered at Baseline and at Week 8. The coefficients of

value 0.99 along the main diagonal indicated that the homonymous factors could be considered equal.

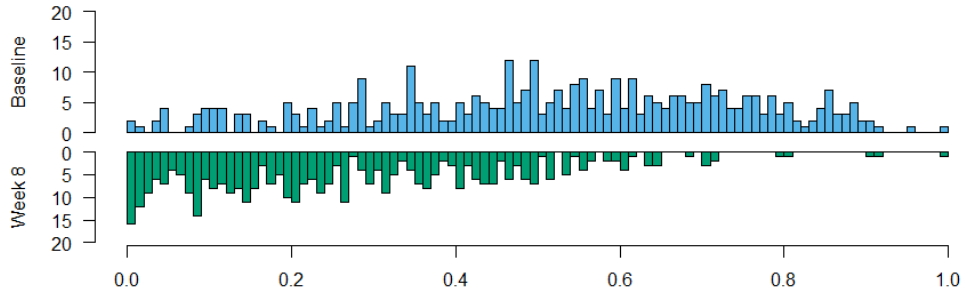
	PRF (W8)	**CF (W8)**	**LF (W8)**
PRF (Baseline)	0.99	0.00	0.00
CF (Baseline)	-0.03	0.99	-0.01
LF (Baseline)	0.01	0.02	0.99

Table 4.1: Tucker's congruence coefficient for the Patient-Reported (PRF), Clinical (CF), and Laboratory (LF) factors uncovered at Baseline and at Week 8 (W8). The coefficients of value 0.99 along the main diagonal indicated that the homonymous factors could be considered equal.

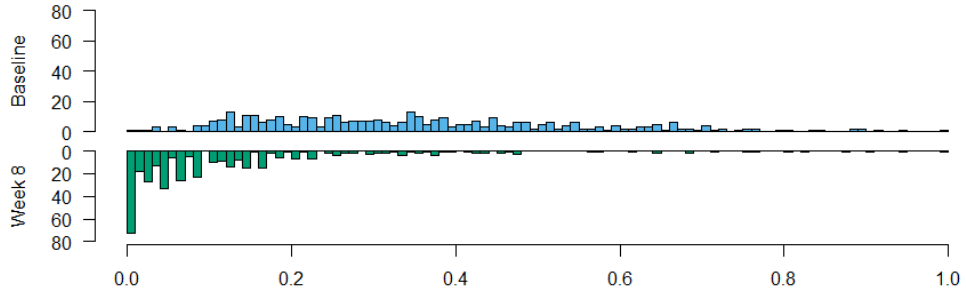
As mentioned in Section 3.3, the factors were scaled to range between 0 and 1. In Figure 4.1 the distributions of the factor scores across the three factors are displayed. The blue histogram (on top) depicted the distribution at Baseline, while the green histogram (bottom) depicted the distribution at Week 8. It is already evident that the values were, on average, smaller at Week 8 than at Baseline: a clear indicator of the efficacy of the treatments. The patient-reported factor at Baseline is the most uniformly distributed, with the scores ranging over the whole interval. Such relatively high scores might be related to the general distress of being diagnosed with a chronic disease. While it is common for RA patients to report residual burden due to unmet needs, at Baseline the patients could also be alarmed and overwhelmed by the diagnosis. This could have potentially led patients to overestimate their perception of the disease, as this might have been the first time they were confronted with it. The higher patient-reported factors could also arise from the general discomfort of the disease activity not being under control at the beginning of the trial. At Week 8, all the distributions displayed a positive skew. The means and standard deviations of the three factor scores for both analyses can be found in Table 4.2. In all three cases, the averages of the factor scores at Week 8 were approximately half or less than the averages of the factor scores at Baseline.

	Baseline	**Week 8**
	mean (SD)	**mean (SD)**
Patient-Reported Factor	0.516 (0.22)	0.266 (0.19)
Clinical Factor	0.354 (0.20)	0.129 (0.17)
Laboratory Factor	0.290 (0.27)	0.147 (0.16)

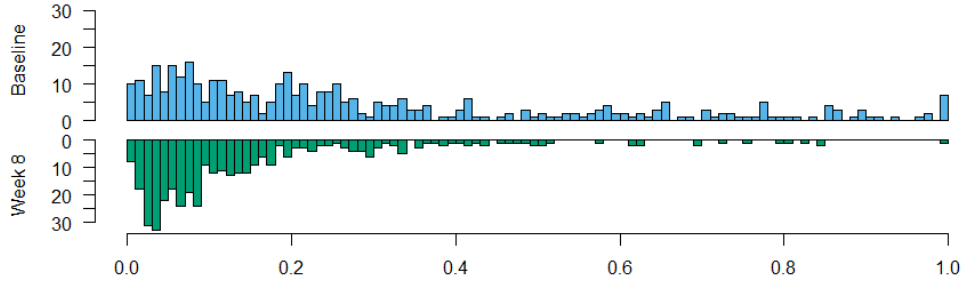
Table 4.2: Means and standard deviations of the three factor scores at Baseline and at Week 8. For all three factors, the means of the factor scores at Week 8 is approximately half the means of the factor scores at Baseline



(a) Patient-Reported Factor



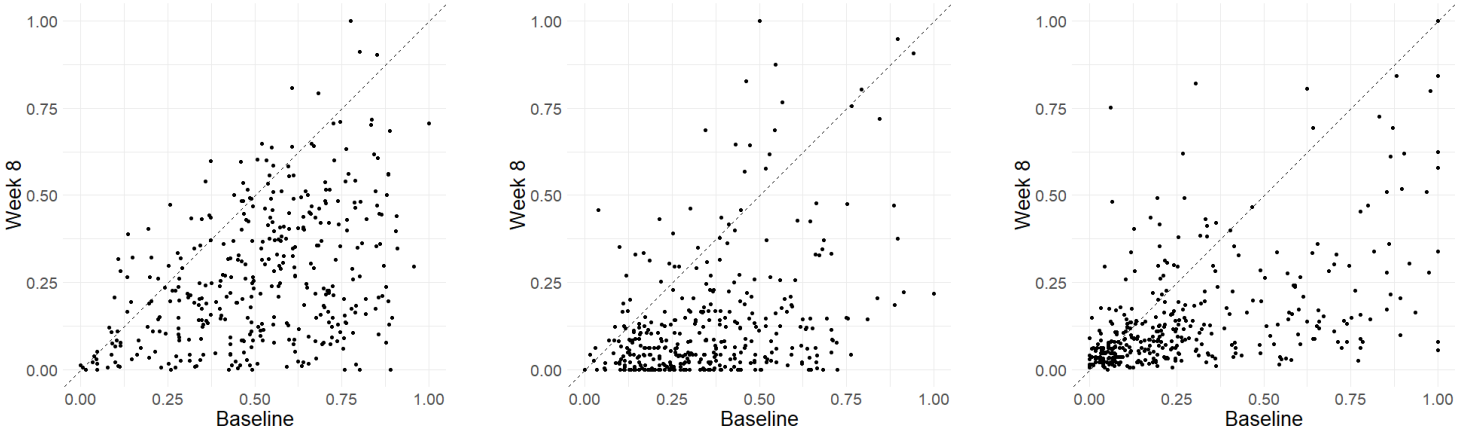
(b) Clinical Factor



(c) Laboratory Factor

Figure 4.1: Distribution of the factor scores for the patient-reported factor (a), clinical factor (b), and laboratory factor (c) for both Baseline (in blue, top) and for Week 8 (in green, bottom). The y-axes of the three sub-figures differ based on the distribution peak.

Figure 4.2 displays the difference of the factor scores for each patient between Baseline and Week 8 for all three factors. The diagonal dotted line represents the line of equality between the two measurements. Thus, any observation falling along this line had equal scores at Baseline and Week 8. Any observation falling below the dotted line indicated an improvement in the corresponding factor for a specific patient. As expected, this applied to the majority of the patients. Particularly, 237 (62.5%) patients exhibited an improvement across all factors from Baseline to Week 8. In contrast, only 8 (2%) patients had higher factor scores at Week 8 than at Baseline for all three factors.



(a) Patient-Reported Factor (b) Clinical Factor (c) Laboratory Factor

Figure 4.2: Comparison of the factor scores for each patient between Baseline and Week 8 for the patient-reported factor (a), clinical factor (b), and laboratory factor (c)

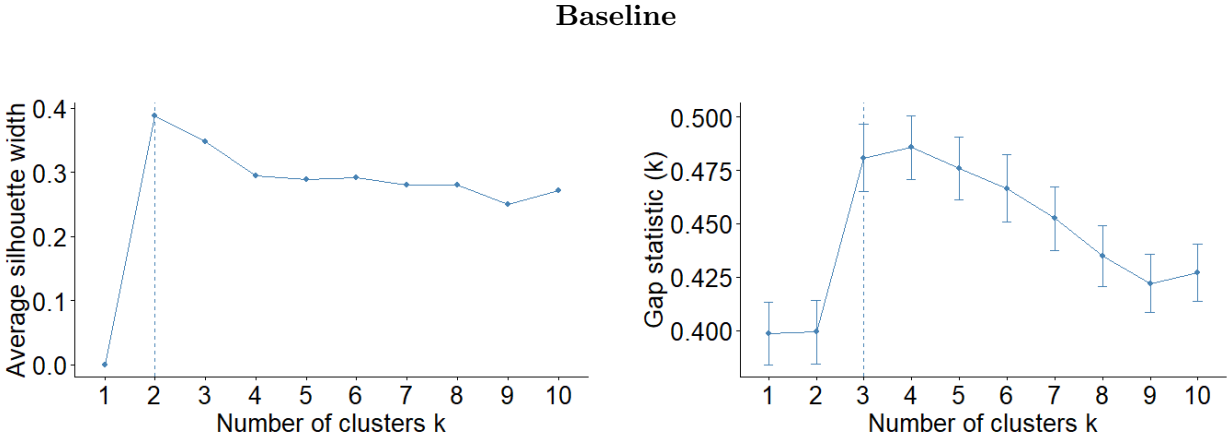
4.2 Clustering

Following the two separate exploratory factor analyses, two cluster analyses were carried out separately: one for the measurements at Baseline and one for the measurements at Week 8.

4.2.1 Number of clusters

The number of clusters was chosen independently in the two analyses. There was no expectation that the optimal number would have to be equal at Baseline and at Week 8. The only numerical constraint was that the number of clusters needed to exceed two. The study population was already split into high-risk and low-risk groups. We were, on the other hand, seeking to differentiate the sample into smaller groups with similar characteristics. Both methods introduced in Section 3.4.2 were employed both at Baseline and at Week 8 for a more educated selection.

- Baseline
  Figure 4.3 displays both the average silhouette and the gap statistic plots for the measurements at Baseline. The former indicated a selection of only two clusters as the optimal choice. As we were looking for larger diversification among clusters, the subsequent optimal choices were three of four clusters. On the other hand, the gap statistic method evaluated a two-cluster solution as sub-optimal when compared to a three or four-cluster one. As specified in the method's description, the gap statistic is inherently predisposed to favour lower results. Thus, the method would identify three as the optimal number, although a four-cluster solution is the one that actually maximizes the gap statistic. A four-cluster solution was conclusively chosen as the preferred one to possibly account for the bias derived from a small sample size. Ultimately, a cluster analysis, like any other unsupervised learning method, is partially subjective. The purpose it served was to produce an outcome that could be meaningful from a clinical perspective. Thus, selecting a cluster solution that was clinically meaningful was also part of the decision making. A four clusters solution yielded results that could be fruitful to rheumatologists. For the selected study population, a four clusters solution provided results that would make the distinction amongst patients and the interpretation of the clusters more straightforward.

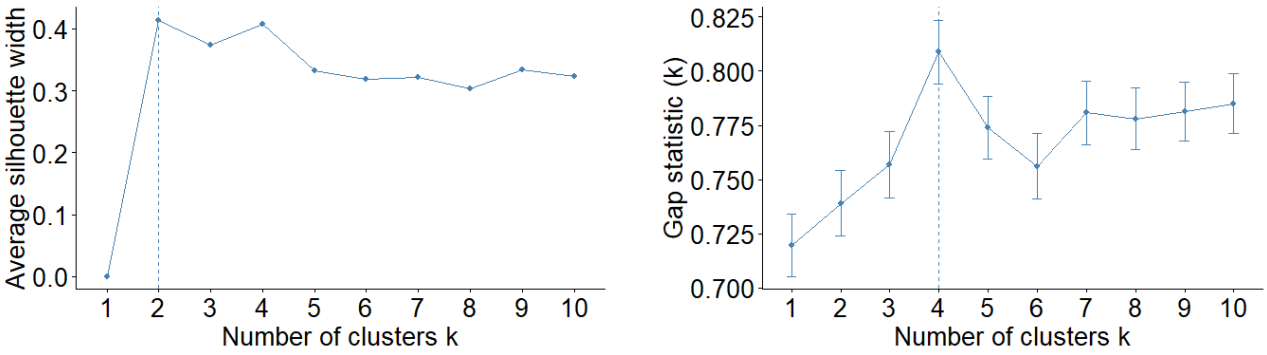


(a) average silhouette method (b) gap statistic method

Figure 4.3: Two different methods were applied at Baseline to find the optimal number of clusters k. The average silhouette method (a) and the gap statistic method (b) were employed for this purpose.

- Week 8
  The analysis at Week 8 was more straightforward and objective. Figure 4.4 displays the two methods for the measurement at Week 8. The average silhouette method indicated either two or four as the optimal number of clusters. The gap statistic method confirms a four-cluster solution as the optimal one. Likewise to the analysis at Baseline, the optimal value of k was set to four.

Week 8



(a) average silhouette method

(b) gap statistic method

Figure 4.4: Two different methods were applied at Week 8 to find the optimal number of clusters k. The average silhouette method (a) and the gap statistic method (b) were employed for this purpose.

4.2.2 Clusters

Hierarchical K-means was performed on both the measurements at Baseline and at Week 8, with the number of clusters k set to four in both cases. The clusters found at Baseline are reported in Table 4.3. The table contains four named clusters, the centroids of their corresponding observations, and the corresponding sizes. The clusters were named after thorough discussion and interpretation of the results with clinical experts.

Baseline

Cluster	**Patient F.** mean (SD)	**Clinical F.** mean (SD)	**Laboratory F.** mean (SD)	**Size**
Mildly symptomatic	0.237 (0.11)	0.220 (0.13)	0.148 (0.11)	107
Unmet needs	0.588 (0.11)	0.260 (0.10)	0.160 (0.12)	120
Non-inflammatory burden	0.700 (0.15)	0.563 (0.13)	0.217 (0.12)	71
Vulnerable	0.619 (0.16)	0.485 (0.20)	0.733 (0.15)	81

Table 4.3: Centroids (and standard deviations) for the Patient-Reported, the Clinical, and the Laboratory factors for each of the four clusters found at Baseline. The size of the cluster is also stated. The numbers are colored based on whether they indicate a relatively low factor score (green), or a relatively high one (orange).

The clusters found displayed unique characteristics:

- Mildly symptomatic
  The first cluster exhibited the lowest average factor scores across all factors. Particularly, the patient-reported factor deviated the most from the rest of the clusters, denoting the prevalence of patients who experienced a relatively low disease burden. The cluster was named "mildly symptomatic" as the low factor scores indicated overall lower measurements in the variables considered in the study.

- Unmet needs
  For both the clinical and the laboratory factors, the second cluster did not deviate greatly from the first one. As indicated by both factor scores, this subgroup of patients did not display common disease activity symptoms such as a high joint count or elevated measurements for the biomarkers. On the other hand, the two clusters' patient-reported factors differed considerably. Possibly, this cluster might have comprised all those patients who felt alarmed by the diagnosis of a chronic disease, but who did not yet exhibit significant complaints. At Baseline, this was the cluster with the most patients.

- Non-inflammatory burden
  This cluster of patients was particularly interesting from a clinical perspective. The patients and the doctors expressed similar evaluations of the disease activity, resulting in the highest patient-reported and clinician factors across the four clusters. The laboratory factor, on the other hand, declared relatively low values for the biomarkers. It is worth noting that the clinician evaluation, similar to the patient-reported one, is imperfect and sensitive to bias. There might be many cases where, especially at baseline, the clinician would carry out a more detailed and careful analysis if the patient were to report a high disease burden. Doctors might be inclined to be more attentive to potential swelling and tenderness in the joints of patients with high disease burden. The interaction between the doctor and the patient could cause potential bias in the evaluation.

- Vulnerable
  The last cluster contained patients with high factor scores. The trait that detached this cluster from the others was the large laboratory factor score. Patients in this cluster typically presented biomarkers much higher than the rest of the sample. Furthermore, given its large standard deviations, it is the most widely distributed cluster found at Baseline.

The clusters found at Week 8 are reported in Table 4.4. As in the previous case, the table contains four named clusters, the centroids of their corresponding observations, and the corresponding sizes. The first three clusters exhibited features comparable to the first three clusters from Baseline. Thus, they were given the same names. Generally, as expected, the treatment effect caused all factors' scores to decrease in magnitude for each pair of homonymous clusters. Any notable difference is mentioned in the cluster's description.

The fourth cluster at Week 8, on the other hand, differed greatly from the fourth cluster at Baseline. For this reason, a new label seemed better suited.

Week 8

Cluster	Patient F. mean (SD)	Clinical F. mean (SD)	Laboratory F. mean (SD)	Size
Mildly symptomatic	0.122 (0.08)	0.051 (0.06)	0.090 (0.07)	192
Unmet needs	0.419 (0.11)	0.124 (0.09)	0.112 (0.09)	114
Non-inflammatory burden	0.535 (0.16)	0.547(0.20)	0.186 (0.20)	38
Inflammatory burden	0.272 (0.19)	0.121 (0.11)	0.532 (0.17)	35

Table 4.4: Centroids (and standard deviations) for the Patient-Reported, the Clinical, and the Laboratory factors for each of the four clusters found at Week 8. The size of the cluster is also stated. The numbers are colored based on whether they indicate a relatively low factor score (green), or a relatively high one (orange).

- Mildly symptomatic
  As in the previous analysis, at Week 8 there existed a cluster comprising all the patients that had overall better conditions than the rest of the study population. Although the factor scores were not identical to the ones at Baseline, the same name was chosen as this cluster exhibited again the lowest average factor scores across all factors. The distribution of the cluster's factor scores diminished, establishing it as the cluster with the smallest variation. Furthermore, its size increased from Baseline, encompassing more than half of the patients.

- Unmet needs
  Analogous to the analysis at Baseline, the clinical and laboratory factors were relatively similar in the first two clusters of the analysis from Week 8. Again, the main difference between the first two clusters was given by the higher patient-reported factor for the second cluster. The name was even more appropriate in the analysis at Week 8. These were the patients for whom the treatment strategies succeeded at lowering the disease activity, but still displayed a high psychological burden.

- Non-inflammatory burden
  This cluster comprised again the patients for whom the patient-reported and clinical evaluations did not align with the laboratory results. Similarly to the analysis at Baseline, the cluster's patient-reported and clinical factor scores were the highest across the four clusters.

- Inflammatory burden
  This was the only cluster that differed from any other at Baseline. Patients in this

cluster featured relatively low patient-reported and clinical factors, but exhibited laboratory results that were higher than in any other clusters at Week 8. Many of these patients might have been classified as either "mildly symptomatic" or with "unmet needs" based on their patient-reported and clinical factors. On the other hand, an inflammation might have caused their laboratory evaluations to soar, thus generating a new, much smaller, cluster.

4.2.3 Cluster membership

After having examined the cluster structure at the two time points, we were interested in investigating the cluster membership shift between the two analyses. We expected the majority of the patients to either remain in the same cluster, or to show an improvement by shifting at Week 8 to a cluster with lower average factor scores than the one at Baseline. The cluster membership shifts are presented in Table 4.5. The leftmost column contains the four different clusters from Baseline, while the top row specifies the four clusters from Week 8.

Cluster	Mildly symptomatic	Unmet needs	Non-inflammatory burden	Inflammatory burden
Mildly symptomatic	80	16	4	7
Unmet needs	51	51	10	8
Non-inflammatory burden	31	24	14	2
Vulnerable	30	23	10	18

Table 4.5: Cluster membership for the two cross-sectional analyses. The leftmost column contains the four different clusters from Baseline, while the top row specifies the four clusters from Week 8.

In line with the expectations, at Week 8 many patients remained in the same cluster they were allocated to at Baseline. 81 patients remained in the "mildly symptomatic" cluster, 51 stayed in the "unmet needs" cluster and 14 patients were classified as "non-inflammatory burden" at both analyses. At Week 8, a considerable number of patients shifted to the "mildly symptomatic" cluster. This is consistent with our predictions, as this cluster was the largest. In general, we considered the values below the main diagonal of Table 4.5 as the groups of patients for whom at least one factor score improved across the two analyses. For example, the majority of patients that were classified as "unmet needs" at Baseline and as "mildly symptomatic" at Week 8 might have had a positive response to treatment resulting in a lower patient-reported factor. Similarly, patients originally belonging to the "non-inflammatory burden" cluster might have had an improvement in the conditions of their joints, and consequent lower clinical factor, if they were then classified as "unmet needs" at Week 8. Not all patients shifted to a cluster with lower average factor scores. For example, 16 patients who were initially categorized as "mildly symptomatic", were subsequently reclassified as "unmet needs" at Week 8.

4.2.4 Interpretation

Table 4.6 displays the demographics, the factor scores, and the treatment strategies distributions across the clusters found in the cross-sectional analyses. The table contains the five clusters found at Baseline and at Week 8, with three of the clusters remaining stable across both analyses. The first eight rows of the table display the results of the analysis at Baseline. Each cluster's centroids and the standard deviations for each factor, average age, percentage of women, percentage of patients that were stratified into the low-risk group at screening, and percentage of patients that presented comorbidities at screening are presented. The remaining rows below the horizontal line display the results of the analysis at Week 8. In addition to the variables examined at Baseline, for each treatment strategy, the distribution of patients into different clusters is shown. The first three treatment strategies were only applied to the patients stratified into the high-risk (HR) group, while the latter two were applied to patients in the low-risk (LR) group. The treatment distributions were not shown at Baseline as the medications had not yet been given to the patients. There were various clinically relevant results:

- At Baseline, the different clusters presented different estimates for the average age of the patients. Specifically, the "vulnerable" cluster had a significantly higher mean age (59.6) when compared to the remaining clusters and the overall mean age (53.9). In general, the group of older patients is a specific subgroup of the average population of RA patients. This is commonly known as the elderly onset RA. This group usually differs from the common RA by having inflammatory measurements of higher magnitude and being more clinically active overall. It is also not rare that the patients with elderly onset RA are more equally distributed with regard to their genders. Accordingly, the "vulnerable" cluster was also the cluster with the least disparity in the patients' genders (only 59% were women, against the 69% of the overall study population).

- For the majority of the treatments, the distribution of the patients among the clusters aligned with the overall distribution of the study population. For example, 53% of the patients who received the COBRA avant-garde treatment belonged to the "mildly symptomatic" cluster at Week 8. This did not deviate significantly from the overall study population, where 51% of the patients were part of the same cluster.
  An exception to this has been observed. Specifically, the majority of patients stratified into the low-risk group who received the Tight Step-up treatment belonged to the "non-inflammatory burden" cluster at Week 8. 34% of the treatment group was assigned to this cluster, against a 10% of the overall study population. The percentage of patients in the "mildly symptomatic" cluster was only 28%, much smaller than the average 51%. Tight Step-up was the only slow-acting treatment considered in the study. The rest of the treatments were all considered fast-acting. It could be possible that, at Week 8, the slow-acting treatment had not yet shown substantial efficacy.

- The distribution across the clusters of the patients that were stratified into the low-risk group at screening differed greatly in the two analyses. Specifically, at Week 8, 61% of the patients in the cluster "non-inflammatory burden" had been stratified into the low-risk group. At Baseline, the percentage was only 28%. This result followed the same reasoning as the previous one. Patients stratified into the low-risk group were randomised into two different treatments: COBRA slim and Tight Step-up. Considering that the latter is a slow-acting drug, half of the low-risk patients might not have exhibited significant improvement in the first 8 weeks. Thus, these patients exhibited, on average, higher factor scores compared to the rest of the study population.

- The treatment COBRA slim was administered to patients of both the low-risk and the high-risk groups. Interestingly, their distribution among the clusters differed based on the original risk group. Overall, the high-risk group that had received the COBRA slim treatment followed the overall trend of distribution among the clusters. 54%, 31%, 7% and 11% of these patients were classified as "mildly symptomatic", "unmet needs", "non-inflammatory burden" and "inflammatory burden" respectively, against the 51%, 30%, 10%, and 9% of the overall study population. On the other hand, the low-risk group that received the same treatment diverged from this distribution. Specifically, 20% and 16% of the patients were classified as "unmet needs" and "non-inflammatory burden" respectively.
  In general, the low-risk group was a very heterogeneous group. Due to different cutoffs, some patients might have had high inflammatory results and no erosions, or relatively low biomarker values but with a higher joint count and would still be considered part of the low-risk group. Thus it could be that there was an overlap between these two specific clusters at Week 8.

Cluster	Mildly symptomatic	Unmet needs	Non-inflammatory burden	Vulnerable	Inflammatory burden	Total
Patients, n (%)	107 (28%)	120 (32%)	71 (19%)	81 (21%)		379
Patient F., $\bar{x}$ [SD]	0.237 [0.11]	0.588 [0.11]	0.700 [0.15]	0.612 [0.16]		0.516 [0.22]
Clinical F., $\bar{x}$ [SD]	0.220 [0.13]	0.260 [0.10]	0.563 [0.13]	0.485 [0.20]		0.353 [0.19]
Laboratory F., $\bar{x}$ [SD]	0.148 [0.11]	0.160 [0.12]	0.217 [0.12]	0.734 [0.15]		0.290 [0.27]
Age, $\bar{x}$ [SD]	51.9 [13.0]	51.2 [11.7]	54.8 [13.2]	59.6 [12.9]		53.9 [13.0]
Sex: women	73%	71%	72%	59%		69%
Risk-group: low risk	33%	17%	28%	18%		24%
Comorbidities: yes	15%	20%	23%	21%		19%
Patients, n (%)	192 (51%)	114 (30%)	38 (10%)		35 (9%)	379
Patient F., $\bar{x}$ [SD]	0.122 [0.08]	0.419 [0.11]	0.535 [0.16]		0.272 [0.19]	0.266 [0.19]
Clinical F., $\bar{x}$ [SD]	0.051 [0.06]	0.124 [0.09]	0.647 [0.20]		0.121 [0.11]	0.130 [0.17]
Laboratory F., $\bar{x}$ [SD]	0.090 [0.07]	0.112 [0.09]	0.186 [0.20]		0.532 [0.17]	0.147 [016]
Age, $\bar{x}$ [SD]	53.6 [12.2]	53.6 [14.8]	54.5 [12.3]		55.8 [11.6]	53.9 [13.0]
Sex: women	69%	73%	58%		69%	69%
Risk-group: low risk	19%	21%	61%		20%	24%
Comorbidities: yes	19%	21%	26%		06%	19%
COBRA Classic (HR), n (%)	55 (56%)	31 (31%)	2 (2%)		10 (10%)	98
COBRA avant-garde (HR), n (%)	49 (53%)	28 (30%)	6 (6%)		10 (11%)	93
COBRA slim (HR), n (%)	52 (53%)	31 (31%)	7 (7%)		8 (8%)	98
COBRA slim (LR), n (%)	23 (53%)	9 (20%)	7 (16%)		4 (9%)	43
Tight Step-up (LR), n (%)	13 (28%)	15 (32%)	16 (34%)		3 (6%)	47

Table 4.6: The table contains the five clusters found at Baseline and at Week 8, of which three remained consistent for the two analyses. The cluster "vulnerable" was only found at Baseline, while the cluster "inflammatory burden" was only found at Week 8. The first eight rows display the results of the analysis at Baseline. Each cluster's centroids and the standard deviations for each factor, average age, percentage of women, percentage of patients that were stratified into the low-risk group at screening, and percentage of patients that presented comorbidities at screening are presented. The remaining rows below the horizontal line display the results of the analysis at Week 8. In addition to the variables examined at Baseline, for each treatment strategy, the distribution of patients into the different clusters is shown. The first three treatment strategies were applied only to those patients stratified into the high-risk (HR) group, while the latter two were applied to patients in the low-risk (LR) group. Data are presented as either mean and standard deviations, or as percentages.

Chapter 5

Methods: Time-incorporated study

The second study was built on the grounds of the results from the cross-sectional analyses. When considering the whole study population, Pazmino et al. detected a swift improvement in the three factors within the initial 8 weeks of the clinical trial. In the following visits, the factors remained overall stable. Consequently, for the second study, the effect of time was ignored. A time-incorporated cluster analysis was carried out, allowing us to uncover all possible clusters that arose from the dataset.

After Week 8, patients were scheduled for eight more visits to monitor their health in the two years following the start of the treatment. The visits were scheduled at Week 16, 28, 40, 52, 65, 78, 91, and 104. To avoid redundancy with the data used in the cross-sectional analyses, the current study employed data collected during the visits occurring after Week 8. When considering all the visits following Week 8, 17% of the values were missing from the dataset; that is 17% of the total of 3032 cells (379 patients $\times$ 8 visits = 3032) had missing data. Week 91 was the visit with the lowest attendance, with 35% of the patients not attending the scheduled visit. When compared to the cross-sectional study, the higher percentage of missing data required adaptation in the methods employed to account for missingness.

The following study was designed to be a proof-of-concept, rather than a formal analysis. The primary goal of this exploratory study is to assess the effectiveness of the methods. Given the relatively small size of the study population considered, it was not feasible to split the patients into different samples to obtain formal validation. The outcomes of the study and the clinical interpretation of the results provide a clear image of the selected dataset, but may not apply to other populations. A series of decisions were reached based on the specific traits of the study population, making the results dependent on these decisions. The limitations of the study are outlined in the current and the next chapters. The data employed was collected originally to investigate the disparities between the treatment strategies. Furthermore, the exploratory factor analysis had already been carried out on the same dataset. A formal validation would be required on a new dataset. Once the results are validated, the methods could be applied to larger and more heterogeneous datasets. The subsequent results could be useful for obtaining cut-offs for the factor scores, thus providing a "traffic light system" that aids rheumatologists in their decision-making. Results that are generalizable to other patient populations would be desirable for the clinical team that provided the data.

Similar to the previous study, the overall objective of this analysis is to identify coherent clusters within the study population. Hence, many of the methods discussed in this chapter have already been introduced in Chapter 3. They will not be illustrated in detail again. Nevertheless, any differences in the application or arguments of the models will be specified.

5.1 Multiple imputation

Similar to the previous study, multiple imputation (MI) allowed us to perform subsequent analyses without resulting in loss of information from the deletion of incomplete observations [45]. MI accounts for the uncertainty of the imputed values by generating M complete datasets to be employed for further analyses. As before, multiple imputation was implemented by chained equations (MICE) adopting classification and regression trees (CART) as conditional models [47], [48].

To obtain reliable imputation, the data from Baseline and Week 8, as well as the six additional variables collected at screening, were employed. Thus, we obtained a matrix with 379 patients and 96 variables (the nine variables measures at Baseline, Week 8, and the eight visits analysed in the current study, plus the six variables collected at screening). The imputation model followed the algorithm presented in Section 3.1. Most of the arguments for the model were kept equal to the values chosen in the previous study. Considering the percentage of missing data was much higher, a larger number of imputations was preferred. Optimally, the number of imputed sets should be at least equal to the missing data percentage [14], [63]. Given an overall percentage of missing data equal to 17%, the number of imputations was set to 20, thus obtaining 20 imputed and complete datasets.

5.2 Multiple outputation

Once the complete datasets were obtained, the variables employed for more reliable imputations were disregarded. Each of the imputed datasets comprised the nine variables measured at the eight visits. Exploratory factor analysis (EFA) could not yet be performed on the datasets, which contained the same variables measured at different time points. A common approach for obtaining samples with independent observations is through multiple outputation (MO), also referred to as within-cluster resampling [64], [65]. This procedure randomly selects a single data point from each cluster of mutually dependent data and applies subsequent statistical analysis on the independent data. In this case, MO randomly selected one visit for each patient. The resulting dataset comprised, for each patient, the nine variables measured at one of the eight visits. The study is labeled as "time-incorporated", since within one dataset there can be measurements of patients assessed at different time points. This process resulted in a subset where all the observations were mutually independent [65]. In order to minimize loss of information and to account for heterogeneity in the datasets, the process was iterated 100 times on each of the 20 imputed datasets, obtaining 2000 datasets [14].

5.3 Factor scores

As in the previous study, EFA was employed to both reduce the number of variables and to uncover latent factors underlying the observed variables. EFA aids in the interpretation of the factors and provides a clearer description when compared to other dimension reduction

techniques. EFA with principal component extraction was applied to the average correlation matrix of the 2000 datasets, employing Rubin's rule for multiple imputation [56]. Following the results of the study performed by Pazmino et al. we decided to extract three factors [14]. Furthermore, there was no reason to assume independence of the factors so an oblique rotation for the factor analysis was chosen. Again, the promax rotation delivered a matrix with the highest primary loadings and was thus chosen as the matrix rotation technique.

The three factors extracted were equivalent to the ones described by Pazmino et al. [14]. These were the patient-reported (PRF), clinical (CF), and laboratory factors (LF). As the factor loadings are necessary for the subsequent explanation of the factor scores, these are already reported in Table 5.1. The factor loadings exceeding the 0.3 threshold are displayed in bold. Further discussion on the factor loadings' description and interpretation is deferred to Chapter 6.

Variables	Factor 1: Patient	Factor 2: Clinical	Factor 3: Laboratory
PaGH	**0.94**	0.01	-0.01
Pain	**0.92**	0.03	-0.02
Fatigue	**0.91**	-0.17	-0.02
HAQ	**0.62**	0.12	0.10
SJC28	-0.19	**0.94**	0.05
TJC28	0.04	**0.85**	-0.05
PhGH	0.20	**0.73**	-0.02
ESR	-0.02	-0.01	**0.86**
CRP	0.00	0.00	**0.85**

Table 5.1: Factor loadings of the three factors (correlation between the observed variable and the latent factor) for the time-incorporated analysis. Displayed in bold are the loadings exceeding the 0.3 threshold. The factors are ordered by the percentage of variance explained.

In each dataset, the variables were scaled so that each variable would range between 0 and 100. The same procedure described in Section 3.3 was carried out. The variables measured on the VAS were unaltered. The HAQ and the two variables measuring the joint count (SJC28 and TJC28) ranged between [0-3] and [0-28] respectively before the transformation. Finally, as in the previous study, the CRP and ESR were respectively scaled from a range of [0-20] and [0-88] to the desired range. Any observation of the CRP or ESR that exceeded the established upper limits was automatically rescaled to 100.

In the previous study, the factor loadings were extracted twice for the analyses at Baseline and at Week 8, obtaining two matrices. In the current analysis, the loadings were extracted

just once from the average correlation matrix, obtaining one set of factor loadings. Thus, the loadings were used globally as weights for each dataset [1].

To compute the factor scores for each patient in each dataset, the scaled variables were multiplied by their factor loadings and summed up so that each factor would comprise the variables that loaded onto it. Employing the results from Table 5.1 the three-factor scores were computed in the following way:

$$PRF = (0.94 * PaGH) + (0.92 * Pain) + (0.91 * Fatigue) + (0.62 * HAQ)$$

$$CF = (0.94 * SJC28) + (0.85 * TJC28) + (0.73 * PhGH)$$

$$LF = (0.86 * CRP) + (0.85 * ESR)$$

Due to the different number of variables for each factor, each factor score was standardized to a [0–1] scale (with higher values indicating a greater health impact). Consequently, each patient had three factor scores ranging between 0 and 1 [23]. As an outcome, a list of 2000 datasets containing the factor scores for each patient was obtained.

5.4 Clustering

We hypothesized that there existed a time-independent clustering of patients, within the time frame of the clinical trial. Such grouping could describe the patient's disease activity throughout the first two years after the beginning of their treatment. A large number of datasets was created when implementing MI and MO. We did not expect the clusters to remain unvaried with respect to the patients' memberships throughout every dataset. On the other hand, this analysis allowed us to identify all the clusters that arose from this study population. Furthermore, by repeating the cluster analysis over a high number of datasets, we could differentiate between the patients who had predominantly stable cluster memberships and patients whose fluctuating factor scores ensued in varying cluster memberships. The identification of patients belonging to very different clusters could help the rheumatologists as well. Greater focus could be directed toward investigating treatment adherence, comorbidities, and unmet needs in patients not predominantly associated with a specific cluster.

5.4.1 Number of clusters

The squared distance was kept as a measure to determine the distance between two patients, and hierarchical K-means was maintained as the algorithm to cluster the patients. On the other hand, a different approach was employed to assess the optimal number of clusters k. In the cross-sectional study, k was determined separately for each of the two time points, based on a separate dataset. In the current analysis, on the other hand, a universal optimal value had to be selected from the collection of all 2000 datasets.

[1] Extracting the factor loadings for each dataset individually would have been possible, although it would have required a high computational time. We decided to not prioritize this process since the factor loadings are in line with the ones extracted by Pazmino et al. [14].

A criterion called CritCF, developed to integrate MI into cluster analysis, was employed. This criterion evaluates partitions according to different numbers of clusters and different quantities of variables. Furthermore, it allows the user to present results such that they reflect the uncertainty related to the imputed values [66]. As previously stated, there is no universal solution to the problem of identifying the optimal number of clusters in a dataset, and this criterion is no exception. The search strategy will not cover all possibilities, and there is no assurance of reaching the global optimum.

The method aims to minimize the within-cluster variance while simultaneously maximizing the between-cluster partition [67]. The CritCF of a partition with k clusters is determined by the formula

$$CritCF = \left(\frac{2m}{2m+1} \cdot \frac{1}{1+W/B} \right)^{\frac{1+log_2(k+1)}{1+log_2(m+1)}}$$

where m represents the number of variables in the dataset, which in our case are the three factors. W and B are, respectively, the within and between-cluster inertias. The within-cluster inertia is the sum of squared distances between each data point and its assigned cluster center. On the other hand, the between-cluster inertia is the sum of squared distances between cluster centers and the overall centroid of the dataset [67]. The within-cluster inertia aims to measure the compactness of the clusters, while between-cluster inertia assesses the separation between clusters. Lastly, k is the number of clusters. Higher values of CritCF indicate a better partition [2].

The CritCF method requires the user to specify a range of number of clusters to be explored. Since both the analyses at Baseline and Week 8 indicated a four-cluster solution to be optimal for the respective dataset, the minimum number of clusters investigated was set to four. The maximum number of clusters to explore was chosen more arbitrarily. The cluster analysis aimed to produce an outcome that could be meaningful from a clinical perspective. When generalized to larger and more heterogeneous datasets, a possible product of the analysis would be the obtainment of cut-offs for the factor scores. A higher number of clusters, and thus a greater quantity of cut-offs, would not be practical for rheumatologists in their decision-making. Therefore, the maximum number of clusters explored was set to eight.

The criterion was measured for values of k ranging between four and eight. Each dataset returned an optimal number of clusters. The final number of clusters was selected as the one emerging with the highest frequency in the 2000 datasets [66]. A four-cluster solution proved to be optimal in most of the datasets. Further details on the choice of the optimal number of clusters are deferred to Chapter 6.

[2] The criterion can also be employed to select an optimal set of variables for the cluster analysis. Variables that, when removed, produce higher CritCF values are excluded from the model. In general, this feature proves useful in high-dimensional datasets where not all variables might be fruitful for the cluster analysis. Variable selection was not implemented in the current study as all three variables were necessary for clustering. Additional information can be found in the paper by Basagaña et al. [66]

5.4.2 Clusters

Through hierarchical K-means clustering, four clusters were created in each of the 2000 datasets. The application of both MI and MO led to a collection of diverse datasets. Thus, there was no expectation that the clusters had to remain unvaried among each dataset, nor that the patients could not shift clusters in different datasets. Nonetheless, we were interested in aligning the clusters from all the datasets based on the vicinity of their centroids. Thus, we could assess the stability of specific clusters in relation to the factors that defined them. This would avoid any potential issues arising from the automatic labeling of the statistical software. If, for example, a cluster was labelled as "first cluster" in one dataset but "second cluster" in the next one it could hinder the correct interpretation of the clusters [66].

A cluster analysis was performed on the 8000 centroids obtained from the four cluster partitions applied to the 2000 datasets. Four global centroids were achieved, representing the centers of the clusters from the datasets. The clusters of each dataset were aligned based on their vicinity to the global centroids, allowing us to identify which clusters were stable across the datasets.

In the previous study, computing the cluster membership was a straightforward process. The cluster analysis provided automatically the allocation of each patient in one of the clusters. In the current study, on the other hand, the results of 2000 cluster analyses needed to be combined. Various metrics were computed to obtain a comprehensive understanding of the cluster membership and size.

- Average cluster size
  The size of the homonymous clusters was computed across all datasets. The average of this measure provided insight into the relative sizes of the clusters across the datasets.

- Cluster assignment by majority vote
  Ultimately, the allocation of each patient to a specific cluster was made by majority vote. Each patient was assigned to the cluster to which they had been assigned most frequently in the 2000 datasets [66]. This procedure yielded a second estimate of the cluster sizes, determined by the number of patients that were assigned to that specific cluster by majority vote.

- Frequency of allocation
  For each patient, we computed the frequency of allocation to the cluster to which they had been assigned by majority vote. Thus, we could inspect patients for whom the cluster membership remained overall stable. Similarly, patients whose fluctuating factors resulted in a continuous shift between clusters were also identified. These patients could be examined further and might be particularly relevant from a clinical perspective. After computing the frequency of allocation for each patient, we determined the average frequency of allocation to the patients' primary clusters. Thus we estimated on average how often patients were allocated to their primary cluster.

Chapter 6

Results and clinical interpretation: Time-incorporated study

Multiple imputation was employed to obtain 20 complete datasets, which could account for the uncertainty deriving from the high percentage of missing data. To achieve datasets with independent observation, multiple outputation was applied 100 times to each of the 20 datasets. A collection of 2000 datasets was obtained, with each dataset comprising mutually independent data.

6.1 Factors

A correlation matrix was computed for each dataset. Through Rubin's rule for multiple imputation, an exploratory factor analysis was performed on the average correlation matrix of the 2000 datasets. Table 5.1 displays the factor loadings extracted from the average correlation matrix. The nomenclature for the factors was maintained consistent with the one employed by Pazmino et al. and our previous study [14]. The three-factor model explained 74% of the variance of the disease activity measured by the nine variables included. The factors were ordered by the percentage of variance explained, with the patient-reported, the clinical, and the laboratory factors explaining 33%, 24%, and 16% of the variance respectively. The variables were ordered by their correlation with the primary loading. Subsequently, the factor scores were obtained for each patient following the equations from Section 5.3. The factor scores were then scaled to range between 0 and 1. Table 6.1 displays the means and standard deviations of the three factor scores across all the datasets.

	mean (SD)
Patient-Reported Factor	0.297 (0.23)
Clinical Factor	0.113 (0.16)
Laboratory Factor	0.163 (0.18)

Table 6.1: Means and standard deviations of the three factor scores across all the datasets.

6.2 Clustering

A cluster analysis was performed separately on each of the 2000 datasets.

6.2.1 Number of clusters

A list of 2000 values, representing the optimal number of clusters for each dataset, was obtained. Each value indicated the outcome generated by the application of CritCF to a separate dataset. Table 6.2 displays, for each value of k considered, the number of datasets for which k would have been selected as the optimal number of clusters. The overall percentage of datasets is also indicated. The global number of clusters, denoted at k_{global}, was selected as the value with the highest occurrence among the 2000 datasets [66].

k	4	5	6	7	8
Occurrence (%)	965 (48%)	340 (17%)	282 (14%)	205 (10%)	208 (11%)

Table 6.2: The occurrence of each value of k within the range [4-8] is presented. This indicates the number of datasets for which that specific k was identified as the optimal number of clusters. The percentage of occurrence of the 2000 datasets is also shown.

An alternative method employed for selecting the number of clusters was to examine the distribution of CritCF across the 2000 datasets based on the number of clusters. The number of clusters exhibiting the highest median CritCF could be considered as a confirmatory technique for the choice of the optimal number of clusters. Figure 6.1 compares the performance of the clustering method for the chosen range of k across all datasets. For the selected study population, this method validates a k_{global} equal to four. However, extracting direct information regarding the percentage of times one would choose a particular k over another is challenging from this type of figure. Commonly, though not obligatorily, a substantial overlap between two box plots suggests that both k values outperform each other in a significant number of datasets [66].

The CritCF criterion was designed to integrate multiple imputation into cluster analysis in a way that the results could reflect the uncertainty arising from the imputed values for the missing observations [66]. The subsequent application of multiple outputation to the imputed datasets resulted in increased heterogeneity among the datasets. This could be one of the reasons for the absence of a universally established optimal number of clusters.

Ultimately, for the selected study population, a four-cluster solution was chosen as it provided a clearer interpretation of the clusters in the majority of the datasets. It should be noted that the selected study population was rather homogeneous from a clinical perspective, as all patients were diagnosed with RA within a year before the start of the trial. It could be likely that a more diversified population may require a higher number of clusters when applying the same methodology.

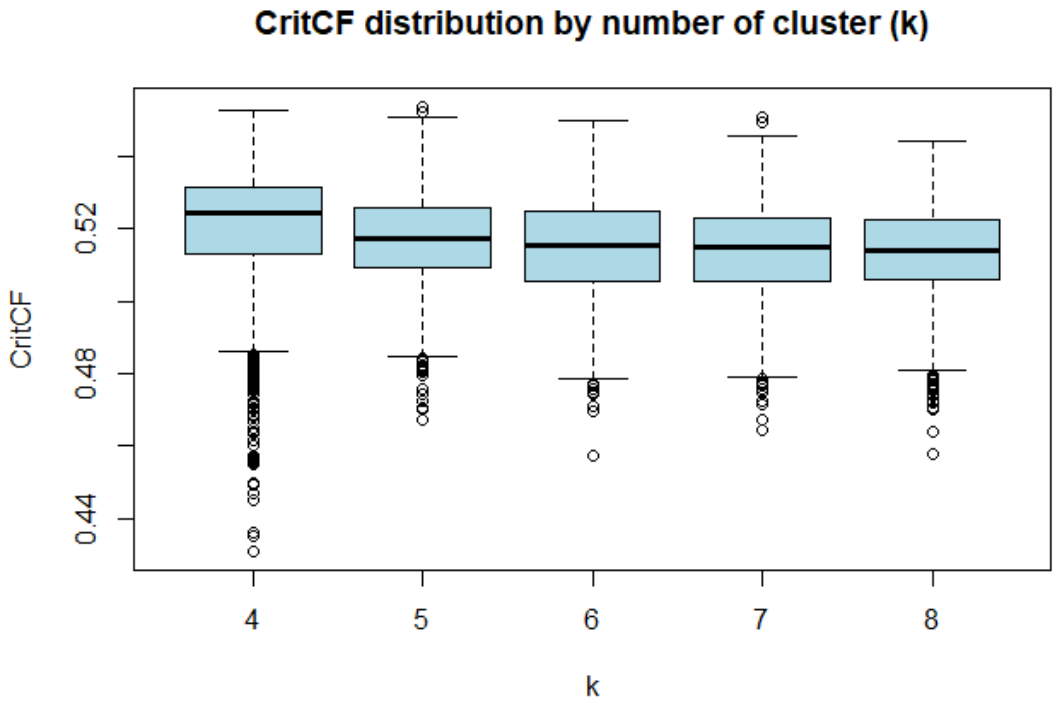


Figure 6.1: Box plots illustrating the between imputations distribution of CritCF based on the number of clusters (k). Each box plot represents the distribution of the CritCF values across the 2000 datasets if that specific k was chosen as the number of clusters [66].

6.2.2 Clusters

Hierarchical K-means clustering was performed separately on each of the 2000 datasets with the number of clusters k set to four. This resulted in 8000 clusters with 8000 centroids. We were interested in aligning the clusters from all the datasets based on the vicinity of their centroids. To obtain a "gold standard" to which each dataset's clusters could be compared, a subsequent hierarchical K-means cluster analysis was performed on the 8000 centroids. The four clusters of each dataset were aligned based on their vicinity to the four global centroids.

Once the clusters have been aligned, the average factor scores for each group of clusters were computed. Table 6.3 reports four named clusters and, for each factor, the average of the centroids of the homonymous clusters across the 2000 datasets.

Cluster	Patient F. mean (SD)	Clinical F. mean (SD)	Laboratory F. mean (SD)
Mildly symptomatic	0.126 (0.02)	0.052 (0.01)	0.108 (0.02)
Unmet needs	0.458 (0.10)	0.113 (0.03)	0.126 (0.04)
Inflammatory burden	0.363(0.11)	0.144 (0.06)	0.596 (0.11)
"Other"	0.639(0.07)	0.424 (0.16)	0.246 (0.13)

Table 6.3: Centroids (and standard deviations) for the Patient-Reported, the Clinical, and the Laboratory factors for each of the four clusters. The estimates were computed from the location of the centroids of the 2000 homonymous clusters in the 2000 datasets. Thus, the standard deviations do not directly represent the variability among the patients grouped within a certain cluster. Instead, they capture the variance inherent to the particular cluster for three factors. The numbers are colored based on whether they indicate a relatively low factor score (green), or a relatively high one (orange).

It should be emphasized that the centroids and the standard deviations in Table 6.3 were computed from the location of the centroids of the 2000 homonymous clusters in the 2000 datasets. Thus, the standard deviations do not directly represent the variability among the patients grouped within a certain cluster. Instead, they capture the variance inherent to the particular cluster for the three factors. The clusters found displayed unique characteristics:

- Mildly symptomatic
  The first cluster was characterized by both the lowest means and standard deviations, when compared to the other clusters, across all factors. Particularly, the patient-reported factor deviated the most from the rest of the clusters, denoting the prevalence of patients who reacted positively to the treatment and experienced a relatively low disease burden. The cluster was assigned the same label as the first cluster from the previous study as it exhibited similar features. The low factor scores indicated overall lower measurements in the variables considered in the study. The low standard deviations denoted stability across the datasets and we hypothesized that patients' membership to this cluster was also overall stable. To clarify, we conjectured that, regardless of the dataset under consideration, the assignment of patients to this particular cluster remained overall consistent.

- Unmet needs
  For both the clinical and the laboratory factors, the second cluster did not deviate greatly from the first one. Both the centroids and the standard deviations were comparable. As indicated by both factor scores, this subgroup of patients did not display common disease activity symptoms such as a high joint count or elevated measurements for the biomarkers. On the other hand, the first two clusters' patient-reported factors

differed considerably. These were the patients for whom the treatment strategies succeeded at lowering the disease activity, but still displayed a relatively high psychological burden. Similarly to the previous study, the cluster was then denoted as "unmet needs". The high patient-reported factor's standard deviation denoted variability in the scores of the patients assigned to this cluster. We hypothesized that some patients could shift membership between the "unmet needs" and the "mildly symptomatic" cluster between different datasets.

- Inflammatory burden
  The clusters classified as "inflammatory burden" exhibited relatively low patient-reported and clinical factors, but featured laboratory results that were higher than in any other cluster. Similarly to the analysis at Week 8, we hypothesized that this cluster would comprise the few patients for whom an inflammation might have caused the laboratory evaluations to soar.

- "Other"
  The fourth cluster exhibited very high standard deviations when compared to the other clusters. Both the patient-reported and the clinical factors were the highest. The high patient-reported mean, in conjunction with its standard deviation, denoted that the patients that have been classified in this cluster across the datasets had a high psychological burden. Given the instability of the cluster concerning the factor scores, we hypothesized that the patients' memberships would also be inconsistent across all datasets. Furthermore, given the lack of a universally definite optimal number of clusters, this cluster could be the one that, in some of the datasets, would have been split into multiple subgroups. In view of these considerations, we did not wish to categorize the patients with a fixed cluster name, as the cluster's factor scores vary greatly across the datasets. Thus, the label "Other" was assigned to the last cluster.

6.2.3 Cluster membership

The average cluster size was determined for each aligned cluster. Patients were ultimately assigned to a cluster by majority vote. Thus, each patient was allocated to the cluster to which they had been assigned most frequently in the 2000 datasets. Nonetheless, the frequency of allocation to every cluster was recorded. The outcomes of these evaluations can be found in Table 6.4.

The proportions of the four clusters remained consistent across the various measures. Together, the "mildly symptomatic" and the "unmet needs" clusters accounted for 80 to 91% of the patients, depending on the criteria employed. The "mildly symptomatic" cluster included the majority of the patients in either evaluation, which could be interpreted as a sign of the efficacy of the overall treatment strategies. As expected, the "inflammatory burden" cluster has a relatively small size. As previously mentioned, the cluster may consist of a small number of patients for whom the elevated laboratory assessments might have been due to chronic inflammation.

On the third row of Table 6.4, the average frequency of allocation to the patients' primary clusters is displayed. A greater value indicates a higher percentage of allocation to the selected clusters across all datasets. On average a patient allocated to the "mildly symptomatic" cluster by majority vote has been classified into the "mildly symptomatic" cluster in 74.6% of the datasets. On the other hand, patients assigned to the "other" cluster through majority vote have been, on average, categorized into the homonymous cluster in only 51.3% of the datasets. Thus, patients allocated by majority vote in the "mildly symptomatic" cluster exhibited a more stable cluster membership across the 2000 datasets than the patients assigned to the other clusters. This result was in line with our expectations. Given the low standard deviations of the first cluster from Table 6.3, we expected a stable membership to this cluster. In opposition, the lower frequency of allocation in the "other" cluster may stem from the higher standard deviations in the factor scores.

6.2.4 Interpretation

Table 6.4 displays the clusters' factor scores across the clusters found in the time-incorporated analysis. For each factor, the clusters' centroids and standard deviations are outlined in two sets: the cluster average and the patient average. The former contains the information presented in Table 6.3 and computes the means and the standard deviations of the centroids of the 2000 homonymous clusters in the 2000 datasets. The patient average factor scores, on the other hand, were computed by averaging all the measurements from each time a patient was allocated to a certain cluster. Thus, this measure reflects the variability among the patients grouped within each cluster. As expected, the centroids' locations do not vary greatly. However, the patient average factor scores have a much higher standard deviation for all factors, representing the variance of the factor scores on a patient level. The cluster where the patient average factor scores deviate mostly from the cluster average factor scores is the one labelled "other", denoting once again the instability of the factor scores and the unfeasibility of finding an appropriate label for the cluster.

The centroids of the factors computed on the total cluster average occurred to be higher than the same measures computed on the patient average. The average cluster centroid for the patient-reported factor, for example, had a value of 0.396. On the other hand, the average of the patient-reported factors across all observations in every dataset was estimated at 0.297. The different magnitude arises from the way each measure was computed. The cluster average combines the values of the 8000 centroids, thus allocating the same weight to clusters of different sizes. In this case, the 2000 clusters labelled as "other" significantly contribute to the overall average increase, despite comprising a small percentage of the patients. Differently, the factor scores average computed directly on the patients provides a better indication of what the point estimates are across the datasets.

The average age, percentage of women, percentage of patients that were stratified into the low-risk group at screening, and percentage of patients that presented comorbidities at screening are also presented for every cluster in Table 6.4. These percentages and evaluations were computed for each cluster by considering only the patients that were assigned to the cluster by majority vote. For each treatment strategy, the distribution of patients into different

clusters is also shown. The first three treatment strategies were only applied to the patients stratified into the high-risk (HR) group, while the latter two were applied to patients in the low-risk (LR) group. There were various clinically relevant results:

- 25% of the patients classified in the "other" cluster presented comorbidities at screening, against the average of 19% in the study population. In this context, the presence of comorbidities and the high factor scores characterizing the cluster are not unrelated. Having comorbidities might influence the various measurements and the treatment results [68]. For example, comorbid cardiovascular diseases could result in inflammation and subsequent higher measurements for the biomarkers. Comorbidities also affect the patient's perception of the disease because distinguishing the symptoms may be difficult. Higher patient-reported factor scores are thus expected in patients with comorbidities.

- The factor scores of the "inflammatory burden" cluster are particularly interesting from a clinical perspective. The high laboratory factor score denotes a group of patients with consistently elevated levels of inflammation. The low clinical factor score, on the other hand, seems inconsistent with the other two factor scores. This could originate from an underestimation of the disease activity by the clinician. Patients have a more comprehensive understanding of their disease activity in between visits. The evaluation by the clinician performed at individual time points may not fully reflect the ongoing effect of the disease activity that is experienced and perceived by the patient.

- The distribution of patients across the four clusters differs substantially among the two treatment strategies administered to the low-risk group. 70% of the low-risk patients that have been randomised to the COBRA slim treatment strategy belonged to the "mildly symptomatic" cluster, whereas only 19% was classified in the "unmet needs" cluster. On the other hand, among patients randomised to the Tight Step-up treatment strategy, the distribution into the two clusters is 49% and 38%, respectively. Since Tight Step-up is a slower-acting drug, the results prove the need to avoid a delay in successful treatment, which is in line with previous analyses [69]. A faster control of the disease influences the disease activity from a biological perspective. The risk of joint destruction and functional impairment is lower for patients under faster-acting treatments. Furthermore, the longer wait for treatment response when using slow-acting treatments influences the patient's perception of the disease. The quality of life and the ability to cope with the disease are affected [69].
  These considerations are in line with the concept of a therapeutic window of opportunity. This is the stage of the disease course where the biological processes are less advanced and more reversible. The progression of the disease and the chance to achieve remission are inherently linked to the time frame during which the treatment starts to take effect [70].

Cluster	Mildly symptomatic	Unmet needs	Inflammatory burden	"Other"	Total
Average cluster size, n (%)	196 (52%)	107 (28%)	33 (9%)	43 (11%)	379
Assignment by majority vote, n (%)	227 (60%)	116 (31%)	20 (5%)	16 (4%)	379
Frequency of allocation	74.6%	56.7%	52.0%	51.3%	
Patient F. (cluster average), $\bar{x}$ [SD]	0.126 [0.02]	0.458 [0.10]	0.363 [0.11]	0.639 [0.07]	0.396 [0.20]
Clinical F. (cluster average), $\bar{x}$ [SD]	0.052 [0.01]	0.113 [0.03]	0.144 [0.06]	0.424 [0.16]	0.183 [0.17]
Laboratory F. (cluster average), $\bar{x}$ [SD]	0.108 [0.02]	0.126 [0.04]	0.596 [0.11]	0.246 [0.13]	0.269 [0.21]
Patient F. (patient average), $\bar{x}$ [SD]	0.128 [0.09]	0.454 [0.15]	0.350 [0.22]	0.633 [0.17]	0.297 [0.23]
Clinical F. (patient average), $\bar{x}$ [SD]	0.053 [0.07]	0.112 [0.11]	0.135 [0.16]	0.367 [0.25]	0.113 [0.16]
Laboratory F. (patient average), $\bar{x}$ [SD]	0.108 [0.09]	0.122 [0.10]	0.561 [0.22]	0.208 [0.19]	0.163 [0.18]
Age, $\bar{x}$ [SD]	54.2 [13]	53.6 [13]	56.5 [17]	48.7 [9]	53.9 [13]
Sex: women	70%	67%	75%	69%	69%
Risk-group: low risk	24%	22%	25%	31%	24%
Comorbidities: yes	21%	15%	20%	25%	19%
COBRA Classic (HR), n (%)	57 (58%)	35 (36%)	3 (3%)	3 (3%)	98
COBRA avant-garde (HR), n (%)	61 (66%)	24 (26%)	5 (5%)	3 (3%)	93
COBRA slim (HR), n (%)	57 (58%)	31 (31%)	7 (7%)	3 (3%)	98
COBRA slim (LR), n (%)	30 (70%)	8 (19%)	3 (7%)	2 (4%)	43
Tight Step-up (LR), n (%)	23 (49%)	18 (38%)	3 (6%)	3 (6%)	47

Table 6.4: The table contains the four clusters found through the analysis applied to the longitudinal data from Week 16 to Week 104. The first three rows depict the average size and average membership allocation measures for each of the clusters. For each factor, the clusters' centroids and standard deviations are outlined in two sets: the cluster average and the patient average. These are computed, respectively, by averaging the centroids of the homonymous clusters, or by averaging the factor scores of the patients allocated in the same cluster. For each cluster, the average age, percentage of women, percentage of patients that were stratified into the low-risk group at screening, and percentage of patients that presented comorbidities at screening are presented. For each treatment strategy, the distribution of patients into the different clusters is shown. The first three treatment strategies were applied only to those patients stratified into the high-risk (HR) group, while the latter two were applied to patients in the low-risk (LR) group. These percentages and evaluations were computed for each cluster by considering only the patients that were assigned to the cluster by majority vote. Data are presented as either mean and standard deviations, or as percentages.

Chapter 7

Discussion

Summary of the results. This thesis presents a workflow to cluster patients recently diagnosed with rheumatoid arthritis into groups with shared characteristics and comparable disease activity. The aim of the research was two-fold. First, we were interested in finding a grouping of patients by employing only the observations from the start of the treatment cross-sectionally. Second, we sought to identify a more general, visit-independent classification of patients that could describe the patients' outcome measures throughout the first two years after the beginning of the treatment. In both cases, we uncovered the existence of four clusters after carrying out cluster analysis on the factors obtained through previous research.

In line with the first research question, the cross-sectional analyses aimed to uncover the characteristics that define the clustering of the patients at the start of their treatment. Table 4.6 provides an insight into this subject. At Baseline, the cluster labelled as "vulnerable" comprised patients with a higher mean age and a higher percentage of men compared to the other clusters. This group of older patients is a specific subgroup of the RA population and is characterized by higher inflammatory measurements. In contrast to this group, the other clusters shared similar socio-demographic characteristics and differed mainly in the three factors that measure disease activity, highlighting their prognostic value.

The analysis applied to longitudinal data in the second part of the thesis addressed the second research question. Here, we aimed to identify the combination of factors that would describe the clustering of patients throughout the measurements of the clinical trial once the treatment effects can be observed (typically after Week 8). Table 6.4 confirms the existence of a group of patients who reacted positively to the treatment and experienced a relatively low disease burden. This cluster, named "mildly symptomatic", is characterized by a rapid and sustained response to the treatment. This subgroup has already been observed in other studies and is a very stable subgroup that does not require adaptation in medication in the medium term [71]. The standard deviations of this cluster's factor scores presented in Table 6.3 confirm the stability of the cluster concerning disease activity. The early identification of patients that belong to the "mildly symptomatic" cluster could be beneficial to rheumatologists. Patients reaching sustained remission might not need to visit a specialist in the medium term. The rheumatologists could then consider the possibility of partly delegating the follow-up process to a family practitioner. Both measures of evaluating the cluster size indicate that this cluster comprises the majority of the patients, which is in line with the literature [71].

The second largest cluster found in both studies was labelled "unmet needs" indicating the psychological distress experienced by patients. This group comprises all those patients for whom the treatment strategies succeeded in lowering the disease activity, but still displayed a relatively high psychological burden. A similar percentage of patients with unmet psychological needs has been encountered in other studies as well [72], [73]. As the treatment strategies

alone do not seem sufficient to mitigate the impact of the disease, a multidisciplinary intervention through referral to a psychologist, or mental health nurse, could be beneficial for the patients belonging to this cluster [23].

The study on longitudinal data also confirms the existence of a group of patients experiencing consistently elevated levels of inflammation, which could originate from the presence of comorbidities. For example, obese patients tend to have higher inflammatory markers, independent of their disease activity [74]. Obesity is also associated with increased rates of chronic pain and fatigue, which may explain, at least in part, the high standard deviation for this cluster's patient-reported factor.

Confirming the empirical findings of these patient subgroups through statistical analyses, even though exploratory, is an important contribution of this thesis.

Discussion of statistical methods. Assessing the limitations and consequences of the statistical approaches chosen for the analyses allows for an exploration of alternative methods. In this section, I will examine five methods employed in the analyses and inspect alternative approaches while considering their limitations and implications.

First, before computing the factor scores for each patient, the variables considered in the analysis were scaled to obtain equal ranges. The C-reactive protein (CRP) and the erythrocyte sedimentation rate (ESR), specifically, were scrutinized. The skewed distributions of their values, and the lack of a universal maximum value to employ for rescaling purposes, prompted us to consider an alternative solution. The upper whisker of both distributions was utilized as the upper limit in the rescaling process, thus denoting any observation above that level as an outlier. Since the primary objective was to conduct a cluster analysis, this process ensured an emphasis on distinctions among most of the patients exhibiting low CRP and ESR values, rather than on isolated outliers. When conducting a similar analysis on a different sample, a thorough inspection of the distribution of the variables is recommended. As indicated by the clinical team that provided the data, the skewed distribution exhibited by our sample could arise from an early detection of the disease. A sample with established RA could display a more normal distribution of the values for CRP and ESR and would require no adaptation in the rescaling procedure.

Second, multiple outputation was employed to generate datasets with independent observations. The method has been implemented so that a single visit for each patient was randomly selected. Alternatively, one could design a model that accounts for the fluctuating percentage of missingness at the different time points. Multiple outputation could be adapted so that visits with a lower percentage of missing data would have a higher probability of being selected. While this could operate as an extra measure to account for the uncertainty generated by missing data, it would also lead to increased homogeneity among the datasets, thus affecting the outcomes of the cluster analysis.

Third, in the study on longitudinal data, a singular set of factor loadings was obtained by carrying out an exploratory factor analysis on the average of 2000 correlation matrices. When applying the methodology to a different dataset, one could carry out separate exploratory factor analyses on each imputed dataset, thus obtaining multiple factor analytic results. Subsequently, these could be combined following the rearrangement of factors to maximize Tucker's congruence coefficient [14], [62]. In this step, the factors conveying similar

meanings in separate analyses are merged. Ultimately, in our case, a simpler analysis was preferred because the primary focus of the project was on cluster analysis. Moreover, the factor structure is very stable, and small differences between the loadings would only have a small impact on the overall analysis. Additionally, since the outcomes of a single exploratory factor analysis were in line with the ones extracted by Pazmino et al. there was no need to incorporate multiple factor analytic results [14]. Thus, a single set of factor loadings was sufficient for the ensuing techniques. However, when considering a new sample for which the factor loadings are unknown, it may be preferable to conduct multiple exploratory factor analyses.

Fourth, instead of hierarchical K-means, density-based or distribution-based algorithms, such as Gaussian mixture model, could have been employed to cluster the patients. Based on the distribution of the data, or the absence of outliers, alternative methods can be applied to a different sample.

Lastly, various alternatives were tested in the pursuit of establishing a "gold standard" to which each dataset's clusters could be compared. The "gold standard" would allow the alignment of clusters with the same meaning from different datasets, which was fruitful in the interpretation of the clusters. The clusters found at Week 8 were originally considered as a possible candidate for comparison. However, the heterogeneity of the 2000 datasets made the alignment unfeasible. This may have been partially due to one treatment arm in the clinical trial using a slower-acting drug than the rest (making the week 8 solution imperfect as the "gold standard"). Using a cluster solution derived from just one dataset as the "gold standard" led to an unstable alignment of the clusters.

A recursive alignment of the clusters was tested as well. The method would place the datasets in a random order and start by aligning the clusters of the second dataset based on the Euclidean distance of their centroids to those of the first dataset. The clusters of the third dataset would then be aligned based on the Euclidean distance of their centroids to the average of the centroids of the two previous datasets, and so forth. Such a method assigned a large weight to the first datasets which then affected the overall alignment of the clusters from the remaining datasets. For this reason, it is more suitable when the datasets to be matched are more heterogeneous. Ultimately, a cluster analysis applied to the 8000 centroids determined the "gold standard" to which the clusters of each dataset could be matched.

Strengths and limitations. A strength of the research is the high quality of the data, due to the source being a clinical trial and the meticulous data management during the trial.

The proof-of-concept analysis produced promising initial results, outlining an analysis path for the selected data for which there is no current precedent. Further research can statistically validate our approach and formalize treatment cut-off points determined by routinely collected variables. This would enable practitioners to develop a traffic light system used in designing treatment plans and would ultimately both improve patient comfort and allow practitioners and healthcare facilities to focus on the patients with the highest need.

Furthermore, the analysis uncovers patients for whom cluster membership is variable, suggesting that individuals who have a clear disease activity course may systematically differ from those for whom treatment responses fluctuate. Further research could explore these differences to give practitioners a better intuition into which patients may need close monitoring.

Another strength of the study is the detailed examination of the variable distributions, including the implementation of cut-off points for rescaling the laboratory variables. This process of extraction of cut-off points can be replicated when conducting cluster analysis on a comparable dataset.

The main limitation of the present analysis stems from features of the current sample. The sample size of the study is small, which can threaten generalizability to a broader population. Similarly, the study population is largely homogeneous, with patients being diagnosed within one year before the start of the trial, and originating from a small geographical area. This may limit the extent to which the results can be applied to RA patients who are not well represented by the data analyzed here.

Another limitation is the exploratory nature of the analysis on the longitudinal data. The generalization of the findings beyond the description of the studied sample should be approached with caution.

Finally, it is a limitation that the present analysis utilizes data that were collected in a previous study that employed stratified randomization, which is not the optimal sampling design for the present study.

Future work. The clinical team that provided the data has carried out a continuation of the CareRA trial. The new trial, denoted as CareRA2020 (EudraCT number:2017-004054-41), can serve as validation of the exploratory research. The team plans to employ the blueprint of the thesis project and apply it to the new dataset. Likewise, similar data is available in the Swedish Rheumatology Quality Register. The register comprises a large part of the Swedish RA population with over 26,000 participants as of December 2013 [75]. The application of the methods to a wider and more heterogeneous dataset could determine the cut-offs for the factor scores used in the cluster analysis.

In considering future directions for this project, an interesting development for the project involves the combination of the results from both studies. We hypothesize that the majority of the patients would be clustered in a homonymous cluster at Week 8 and in the analysis performed on the longitudinal data. This concept could be generalized further to identify the earliest time-point at which the patients' factor scores can predict the subsequent allocation to specific clusters.

An additional development of the project would be the employment of other variables in predicting cluster membership. There exists a questionnaire, denoted as RAQoL, tailored to evaluate the quality of life (QoL) specific to RA [76]. The 30-item questionnaire evaluated at baseline can be employed to predict cluster membership in the study applied to longitudinal data. The process could be reversed by predicting the RAQoL measurement at the end of the trial based on the patient's cluster allocation.

Conclusion. Through cluster analysis and the inspection of the allocation frequency to the clusters, it is possible to identify those patients with rapid and sustained responses to the treatment. The referral of these patients to a family practitioner would decrease the workload of the rheumatologists. In contrast, a more holistic treatment strategy can be developed for the patients that have been classified in the "inflammatory burden" and "other" clusters.

Acknowledgments

I would like to express my gratitude to the Rheumatology team led by Professor Verschueren at the Katholieke Universiteit Leuven for providing the data that have been used in this research. Sofia Pazmiño, Michaël Doumen, Elias De Meyst, and Professor P. Verschueren significantly contributed to the outline of this thesis by providing clinical insight and interpretation of the outcomes. The findings have been thoroughly interpreted and discussed, thanks to the valuable insights and feedback received.

I would like to extend my appreciation to my supervisor, Dr. Anikó Lovik, for the support and mentorship received throughout the thesis process. The writing of the thesis has been possible thanks to the constant revision and feedback received on the various drafts.

Reference list

[1] J. Bullock, S. A. A. Rizvi, A. M. Saleh, S. S. Ahmed, D. P. Do, R. A. Ansari, and J. Ahmed, "Rheumatoid arthritis: A brief overview of the treatment," *Medical Principles and Practice*, vol. 27, no. 6, pp. 501–507, 2019. DOI: 10.1159/000493390.

[2] P. L. van Riel and A. M. van Gestel, "Clinical outcome measures in rheumatoid arthritis," *Annals of the Rheumatic Diseases*, vol. 59, no. Suppl 1, pp. i28–i32, 2000. DOI: 10.1136/ard.59.suppl_1.i28.

[3] A. F. Radu and S. G. Bungau, "Management of rheumatoid arthritis: An overview," *Cells*, vol. 10, no. 11, p. 2857, 2021. DOI: 10.3390/cells10112857.

[4] L. Innala, E. Berglin, B. Möller, L. Ljung, T. Smedby, A. Södergren, S. Magnusson, S. Rantapää-Dahlqvist, and S. Wållberg-Jonsson, "Age at onset determines severity and choice of treatment in early rheumatoid arthritis: A prospective study," *Arthritis Research & Therapy*, vol. 16, pp. 1–9, 2014. DOI: 10.1186/ar4540.

[5] K. Laiho, J. Tuomilehto, and R. Tilvis, "Prevalence of rheumatoid arthritis and musculoskeletal diseases in the elderly population," *Rheumatology International*, vol. 20, pp. 85–87, 2001. DOI: 10.1007/s002960000087.

[6] Y. Alamanos and A. A. Drosos, "Epidemiology of adult rheumatoid arthritis," *Autoimmunity Reviews*, vol. 4, no. 3, pp. 130–136, 2005. DOI: 10.1016/j.autrev.2004.09.002.

[7] A. Hresko, T. C. Lin, and D. H. Solomon, "Medical care costs associated with rheumatoid arthritis in the US: A systematic literature review and meta-analysis," *Arthritis Care & Research*, vol. 70, no. 10, pp. 1431–1438, 2018. DOI: 10.1002/acr.23512.

[8] E. Myasoedova, J. M. Davis, C. S. Crowson, and S. E. Gabriel, "Epidemiology of rheumatoid arthritis: Rheumatoid arthritis and mortality," *Current Rheumatology Reports*, vol. 12, pp. 379–385, 2010. DOI: 10.1007/s11926-010-0117-y.

[9] J. A. Aviña-Zubieta, H. K. Choi, M. Sadatsafavi, M. Etminan, J. M. Esdaile, and D. Lacaille, "Risk of cardiovascular mortality in patients with rheumatoid arthritis: A meta-analysis of observational studies," *Arthritis Care & Research*, vol. 59, no. 12, pp. 1690–1697, 2008. DOI: 10.1002/art.24092.

[10] S. Sihvonen, M. Korpela, P. Laippala, J. Mustonen, and A. Pasternack, "Death rates and causes of death in patients with rheumatoid arthritis: A population-based study,"

Scandinavian Journal of Rheumatology, vol. 33, no. 4, pp. 221–227, 2004. DOI: 10.1080/03009740410005845.

[11] M. Ishida, Y. Kuroiwa, E. Yoshida, M. Sato, D. Krupa, N. Henry, K. Ikeda, and Y. Kaneko, "Residual symptoms and disease burden among patients with rheumatoid arthritis in remission or low disease activity: A systematic literature review," *Modern Rheumatology*, vol. 28, no. 5, pp. 789–799, 2018. DOI: 10.1080/14397595.2017.1416940.

[12] F. S. Luyster, E. R. Chasens, M. C. Wasko, and J. Dunbar-Jacob, "Sleep quality and functional disability in patients with rheumatoid arthritis," *Journal of Clinical Sleep Medicine*, vol. 7, no. 1, pp. 49–55, 2011. DOI: 10.5664/jcsm.28041.

[13] M. Margaretten, L. Julian, P. Katz, and E. Yelin, "Depression in patients with rheumatoid arthritis: Description, causes and mechanisms," *International Journal of Clinical Rheumatology*, vol. 6, no. 6, pp. 617–623, 2011.

[14] S. Pazmino, A. Lovik, A. Boonen, D. De Cock, V. Stouten, J. Joly, D. Bertrand, K. Van der Elst, R. Westhovens, and P. Verschueren, "Does including pain, fatigue, and physical function when assessing patients with early rheumatoid arthritis provide a comprehensive picture of disease burden?" *The Journal of Rheumatology*, vol. 48, no. 2, pp. 174–178, 2021. DOI: 10.3899/jrheum.200758.

[15] P. L. van Riel, "The development of the disease activity score (DAS) and the disease activity score using 28 joint counts (DAS28)," *Clinical and Experimental Rheumatology*, vol. 32, no. 5 Suppl 85, S65–S74, 2014.

[16] D. M. van der Heijde, M. A. van 't Hof, P. L. van Riel, L. A. Theunisse, E. W. Lubberts, M. A. van Leeuwen, M. H. van Rijswijk, and L. B. van de Putte, "Judging disease activity in clinical practice in rheumatoid arthritis: First step in the development of a disease activity score," *Annals of the Rheumatic Diseases*, vol. 49, no. 11, pp. 916–920, 1990. DOI: 10.1136/ard.49.11.916.

[17] D. M. van der Heijde, M. A. van't Hof, P. L. van Riel, M. A. van Leeuwen, M. H. van Rijswijk, and L. B. van de Putte, "Validity of single variables and composite indices for measuring disease activity in rheumatoid arthritis," *Annals of the Rheumatic Diseases*, vol. 51, no. 2, pp. 177–181, 1992. DOI: 10.1136/ard.51.2.177.

[18] H. A. Fuchs, R. H. Brooks, L. F. Callahan, and T. Pincus, "A simplified twenty-eight–joint quantitative articular index in rheumatoid arthritis," *Arthritis & Rheumatism*, vol. 32, no. 5, pp. 531–537, 1989. DOI: 10.1002/anr.1780320504.

[19] D. Felson, "Defining remission in rheumatoid arthritis," *Annals of the Rheumatic Diseases*, vol. 71, no. Suppl 2, pp. i86–i88, 2012. DOI: 10.1136/annrheumdis-2011-200618.

[20] J. S. Smolen, D. Aletaha, J. W. Bijlsma, F. C. Breedveld, D. Boumpas, G. Burmester, B. Combe, M. Cutolo, M. de Wit, M. Dougados, *et al.*, "Treating rheumatoid arthritis to target: Recommendations of an international task force," *Annals of the Rheumatic Diseases*, vol. 69, no. 4, pp. 631–637, 2010. DOI: 10.1136/ard.2009.123919.

[21] P. C. Taylor, B. Fautrel, Y. Piette, S. Romero-Yuste, J. Broen, M. Welcker, O. Howell, E. Rottier, M. Zignani, K. Van Beneden, *et al.*, "Treat-to-target in rheumatoid arthritis: A real-world study of the application and impact of treat-to-target within the wider context of patient management, patient centricity and advanced therapy use in Europe," *RMD Open*, vol. 8, no. 2, 2022. DOI: 10.1136/rmdopen-2022-002658.

[22] L. H. van Tuyl, M. Sadlonova, S. Hewlett, B. Davis, C. Flurey, N. Goel, L. Gossec, C. Heegaard Brahe, C. L. Hill, W. Hoogland, *et al.*, "The patient perspective on absence of disease activity in rheumatoid arthritis: A survey to identify key domains of patient-perceived remission," *Annals of the Rheumatic Diseases*, vol. 76, no. 5, pp. 855–861, 2017. DOI: 10.1136/annrheumdis-2016-209835.

[23] S. Pazmino, A. Lovik, A. Boonen, D. De Cock, V. Stouten, J. Joly, M. Doumen, D. Bertrand, R. Westhovens, and P. Verschueren, "New indicator for discordance between patient-reported and traditional disease activity outcomes in patients with early rheumatoid arthritis," *Rheumatology*, vol. 62, no. 1, pp. 108–115, 2023. DOI: 10.1093/rheumatology/keac213.

[24] I. Joliffe and B. Morgan, "Principal component analysis and exploratory factor analysis," *Statistical Methods in Medical Research*, vol. 1, no. 1, pp. 69–95, 1992. DOI: 10.1177/096228029200100105.

[25] P. Verschueren, D. De Cock, L. Corluy, R. Joos, C. Langenaken, V. Taelman, F. Raeman, I. Ravelingien, K. Vandevyvere, J. Lenaerts, *et al.*, "Methotrexate in combination with other DMARDs is not superior to methotrexate alone for remission induction with moderate-to-high-dose glucocorticoid bridging in early rheumatoid arthritis after 16 weeks of treatment: The CareRA trial," *Annals of the Rheumatic Diseases*, vol. 74, no. 1, pp. 27–34, 2015. DOI: 10.1136/annrheumdis-2014-205489.

[26] P. Verschueren, D. De Cock, L. Corluy, R. Joos, C. Langenaken, V. Taelman, F. Raeman, I. Ravelingien, K. Vandevyvere, J. Lenaerts, *et al.*, "Effectiveness of methotrexate with step-down glucocorticoid remission induction (COBRA Slim) versus other intensive treatment strategies for early rheumatoid arthritis in a treat-to-target approach: 1-year results of CareRA, a randomised pragmatic open-label superiority trial," *Annals of the*

Rheumatic Diseases, vol. 76, no. 3, pp. 511–520, 2017. DOI: 10.1136/annrheumdis-2016-209212.

[27] P. Verschueren, D. De Cock, L. Corluy, R. Joos, C. Langenaken, V. Taelman, F. Raeman, I. Ravelingien, K. Vandevyvere, J. Lenaerts, *et al.*, "Patients lacking classical poor prognostic markers might also benefit from a step-down glucocorticoid bridging scheme in early rheumatoid arthritis: Week 16 results from the randomized multicenter CareRA trial," *Arthritis Research & Therapy*, vol. 17, no. 1, 2015.

[28] M. H. Weisman, *Rheumatoid Arthritis*. New York, NY, USA: Oxford University Press, 2011, p. 81.

[29] J. E. Pope and E. H. Choy, "C-reactive protein and implications in rheumatoid arthritis and associated comorbidities," *Seminars in Arthritis and Rheumatism*, vol. 51, no. 1, pp. 219–229, 2021. DOI: 10.1016/j.semarthrit.2020.11.005.

[30] FDA, *Review criteria for assessment of C-reactive protein (CRP), high sensitivity C-reactive protein (hsCRP) and cardiac C-reactive protein (cCRP) assays - guidance for industry and FDA staff*, 2005. [Online]. Available: https://www.fda.gov/media/71337/download (visited on 02/01/2024).

[31] J. S. Smolen and P. E. Lipsky, *Contemporary Targeted Therapies in Rheumatology*. London, England: Informa Healthcare, 2007, pp. 601–616.

[32] T. Sokka and T. Pincus, "Erythrocyte sedimentation rate, C-reactive protein, or rheumatoid factor are normal at presentation in 35%–45% of patients with rheumatoid arthritis seen between 1980 and 2004: Analyses from Finland and the United States," *The Journal of Rheumatology*, vol. 36, no. 7, pp. 1387–1390, 2009. DOI: 10.3899/jrheum.080770.

[33] J. Kay, O. Morgacheva, S. P. Messing, J. M. Kremer, J. D. Greenberg, G. W. Reed, E. M. Gravallese, and D. E. Furst, "Clinical disease activity and acute phase reactant levels are discordant among patients with active rheumatoid arthritis: Acute phase reactant levels contribute separately to predicting outcome at one year," *Arthritis Research & Therapy*, vol. 16, 2014. DOI: 10.1186/ar4469.

[34] C. Saadeh, "The erythrocyte sedimentation rate: Old and new clinical applications," *Southern Medical Journal*, vol. 91, no. 3, pp. 220–225, 1998.

[35] K. D. Pagana, T. J. Pagana, and T. N. Pagana, *Mosby's Diagnostic and Laboratory Test Reference*, 12th ed. St. Louis, MO, USA: Elsevier, 2015, pp. 393–394.

[36] D. Mainland, "A seven-day variability study of 499 patients with peripheral rheumatoid arthritis," *Arthritis & Rheumatology*, vol. 8, no. 2, pp. 302–334, 1965. DOI: 10.1002/art.1780080214.

[37] D. A. Delgado, B. S. Lambert, N. Boutris, P. C. McCulloch, A. B. Robbins, M. R. Moreno, and J. D. Harris, "Validation of digital visual analog scale pain scoring with a traditional paper-based visual analog scale in adults," *JAAOS: Global Research and Reviews*, vol. 2, no. 3, 2018. DOI: 10.5435/jaaosglobal-d-17-00088.

[38] D. De Cock and J. Hirsh, "The rheumatoid arthritis patient global assessment: Improve it or lose it!" *Rheumatology*, vol. 59, no. 5, pp. 923–924, 2020. DOI: 10.1093/rheumatology/kez566.

[39] B. Bruce and J. F. Fries, "The health assessment questionnaire (HAQ)," *Clinical and Experimental Rheumatology*, vol. 23, no. 5, S14–S18, 2005.

[40] C. Swales and C. Bulstrode, *Rheumatology, Orthopaedics and Trauma at a Glance*, 2nd ed. Chichester, England: Wiley-Blackwell, 2012, pp. 52–56.

[41] W. J. Falkenburg, D. van Schaardenburg, P. Ooijevaar-de Heer, G. Wolbink, and T. Rispens, "IgG subclass specificity discriminates restricted IgM rheumatoid factor responses from more mature anti–citrullinated protein antibody–associated or isotype-switched IgA responses," *Arthritis & Rheumatology*, vol. 67, no. 12, pp. 3124–3134, 2015. DOI: 10.1002/art.39299.

[42] G. Schett and E. Gravallese, "Bone erosion in rheumatoid arthritis: Mechanisms, diagnosis and treatment," *Nature Reviews Rheumatology*, vol. 8, no. 11, pp. 656–664, 2012. DOI: 10.1038/nrrheum.2012.153.

[43] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976. DOI: 10.1093/biomet/63.3.581.

[44] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002. DOI: 10.1037/1082-989x.7.2.147.

[45] L. F. Burgette and J. P. Reiter, "Multiple imputation for missing data via sequential regression trees," *American Journal of Epidemiology*, vol. 172, no. 9, pp. 1070–1076, 2010. DOI: 10.1093/aje/kwq260.

[46] G. Molenberghs and M. G. Kenward, "Multiple imputation," in *Missing Data in Clinical Studies*, Chichester, England: Wiley, 2007, pp. 105–117.

[47] S. v. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, 2011. DOI: 10.18637/jss.v045.i03.

[48] B. Ripley, *Tree: Classification and regression trees*, 2009. [Online]. Available: https://cran.r-project.org/web/packages/tree/tree.pdf (visited on 02/01/2024).

[49] A. Lovik, V. Nassiri, G. Verbeke, and G. Molenberghs, "Combining factors from different factor analyses based on factor congruence," in *Quantitative Psychology*, S. Culpepper, J. González, R. Janssen, D. Molenaar, and M. Wiberg, Eds., Cham, Switzerland: Springer International Publishing, 2018, pp. 211–219.

[50] T. Rietveld and R. v. Hout, *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin, Germany: Mouton de Gruyter, 1993, p. 292.

[51] L. L. Thurstone, *Multiple Factor Analysis*. Chicago, IL, USA: University of Chicago Press, 1947, p. 335.

[52] R. L. Gorsuch, *Factor analysis*, 2nd ed. Hillsdale, NJ, USA: L. Erlbaum Associates, 1983.

[53] Y. Li, Z. Wen, K.-T. Hau, K.-H. Yuan, and Y. Peng, "Effects of cross-loadings on determining the number of factors to retain," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 27, no. 6, pp. 841–863, 2020. DOI: 10.1080/10705511.2020.1745075.

[54] A. Field, *Discovering Statistics Using SPSS for Windows: Advanced Techniques for the Beginner*. London, UK: Sage, 2000, pp. 438–439.

[55] V. Nassiri, A. Lovik, G. Molenberghs, and G. Verbeke, "On using multiple imputation for exploratory factor analysis of incomplete data," *Behavior Research Methods*, vol. 50, no. 2, pp. 501–517, 2018. DOI: 10.3758/s13428-017-1000-9.

[56] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York, NY, USA: Wiley, 2004, pp. 15–18.

[57] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2009, pp. 349–399.

[58] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, pp. 501–527.

[59] T. Kodinariya and P. Makwana, "Review on determining of cluster in K-means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 90–95, 2013.

[60] A. Kassambara, *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. Sthda, 2017, pp. 36–79. [Online]. Available: https://xsliulab.github.io/Workshop/2021/week10/r-cluster-book.pdf (visited on 02/01/2024).

[61] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 63, no. 2, pp. 411–423, 2001. DOI: 10.1111/1467-9868.00293.

[62] U. Lorenzo-Seva and J. Berge, "Tucker's congruence coefficient as a meaningful index of factor similarity," *Methodology: European Journal of Research Methods for The Behavioral and Social Sciences*, vol. 2, no. 2, pp. 57–64, 2006. DOI: 10.1027/1614-2241.2.2.57.

[63] T. E. Bodner, "What improves with increased missing data imputations?" *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 15, no. 4, pp. 651–675, 2008. DOI: 10.1080/10705510802339072.

[64] E. B. Hoffman, P. K. Sen, and C. R. Weinberg, "Within-cluster resampling," *Biometrika*, vol. 88, no. 4, pp. 1121–1134, 2001.

[65] D. Follmann, M. Proschan, and E. Leifer, "Multiple outputation: Inference for complex clustered data by averaging analyses from independent data," *Biometrics*, vol. 59, no. 2, pp. 420–429, 2003.

[66] X. Basagaña, J. Barrera-Gómez, M. Benet, J. M. Antó, and J. Garcia-Aymerich, "A framework for multiple imputation in cluster analysis," *American Journal of Epidemiology*, vol. 177, no. 7, pp. 718–725, 2013. DOI: 10.1093/aje/kws289.

[67] M. Breaban and H. Luchian, "A unifying criterion for unsupervised clustering and feature selection," *Pattern Recognition*, vol. 44, no. 4, pp. 854–865, 2011, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2010.10.006.

[68] P. C. Taylor, F. Atzeni, A. Balsa, L. Gossec, U. Müller-Ladner, and J. Pope, "The key comorbidities in patients with rheumatoid arthritis: A narrative review," *Journal of Clinical Medicine*, vol. 10, no. 3, p. 509, 2021. DOI: 10.3390/jcm10030509.

[69] P. Verschueren, G. Esselens, and R. Westhovens, "Daily practice effectiveness of a step-down treatment in comparison with a tight step-up for early rheumatoid arthritis," *Rheumatology*, vol. 47, no. 1, pp. 59–64, 2008. DOI: 10.1093/rheumatology/kem288.

[70] L. E. Burgers, K. Raza, and A. H. Van Der Helm-Van, "Window of opportunity in rheumatoid arthritis - definitions and supporting evidence: From old to new perspectives," *RMD Open*, vol. 5, no. 1, 2019. DOI: 10.1136/rmdopen-2018-000870.

[71] M. N. Lwin, L. Serhal, C. Holroyd, and C. J. Edwards, "Rheumatoid arthritis: The impact of mental health on disease: A narrative review," *Rheumatology and Therapy*, vol. 7, pp. 457–471, 2020. DOI: 10.1007/s40744-020-00217-4.

[72] M. J. Walter, T. Kuijper, J. Hazes, A. Weel, and J. Luime, "Fatigue in early, intensively treated and tight-controlled rheumatoid arthritis patients is frequent and persistent: A prospective study," *Rheumatology International*, vol. 38, no. 9, pp. 1643–1650, 2018. DOI: 10.1007/s00296-018-4102-5.

[73] P. C. Taylor, A. Moore, R. Vasilescu, J. Alvir, and M. Tarallo, "A structured literature review of the burden of illness and unmet needs in patients with rheumatoid arthritis: A current perspective," *Rheumatology International*, vol. 36, no. 5, pp. 685–695, 2016. DOI: 10.1007/s00296-015-3415-x.

[74] D. Poudel, M. D. George, and J. F. Baker, "The impact of obesity on disease activity and treatment response in rheumatoid arthritis," *Current Rheumatology Reports*, vol. 22, 2020. DOI: 10.1007/s11926-020-00933-4.

[75] J. Eriksson, J. Askling, and E. Arkema, "The swedish rheumatology quality register: Optimisation of rheumatic disease assessments using register-enriched data," *Clinical and Experimental Rheumatology*, vol. 32 Suppl 85, pp. 147–149, 2014.

[76] G. Tijhuis, Z. De Jong, A. Zwinderman, W. Zuijderduin, L. Jansen, J. Hazes, and T. Vliet Vlieland, "The validity of the rheumatoid arthritis quality of life (RAQoL) questionnaire," *Rheumatology*, vol. 40, no. 10, pp. 1112–1119, 2001. DOI: 10.1093/rheumatology/40.10.1112.

Appendix A

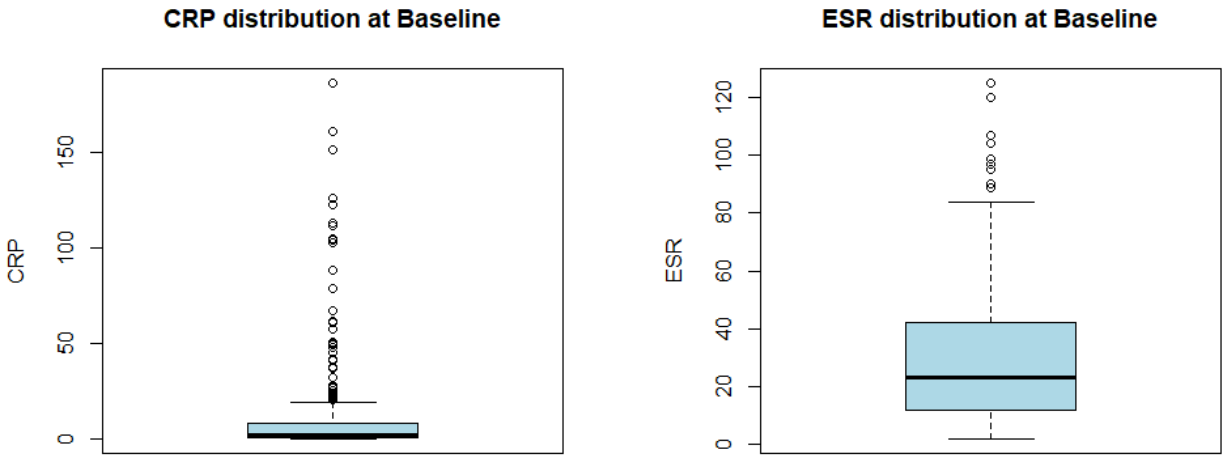
The code is available at: `https://github.com/gabrifresia/Master-Thesis-Stat-DS`. Two files are included, one for the cross-sectional analyses, and one for the time-incorporated analysis. The data is not available to the public.

Appendix B

In Section 3.3, we described the process of rescaling the variables so that each variable would have a range between 0 and 100. As discussed on pages 18 and 19, a specific approach was adopted to rescale the CRP and ESR variables. Since these are biomarkers found in the blood, there is no natural maximum value that can be used for scaling the variables.

Boxplots of the distribution of the two variables at Baseline, brought us to consider an alternative solution. As shown in Figure 1, the upper quartile of the boxplot of the CRP measurements at Baseline had a value of 8.4. This means that 75% of the observations had a value of at most 8.4. At Baseline, the upper whisker (upper quartile times 1.5 times the interquartile range) for the CRP measurement had a value of 20, with 44 patients exhibiting a larger measurement. This number was ultimately chosen as the upper threshold for the scaling.

Similarly, at Baseline, the upper whisker of the boxplot representing the distribution of the ESR values had a value of 88, with just 11 patients displaying a level greater than the threshold. Again, the upper whisker was chosen as the upper threshold for the scaling.



(a) Distribution of the CRP values at Baseline

(b) Distribution of the ESR values at Baseline

Figure 1: Distribution of the biomarkers at Baseline. The features of the boxplots, namely the values of the upper whisker, were employed in the scaling of the variables. Further information regarding the scaling process can be found in Section 3.3