TU WIEN Informatics

Exploring and understanding the impact of machine learning choices on radiotherapy decision making

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Adam Böröndy, Bsc
Matrikelnummer 01610133

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Dr. Renata Raidou
Mitwirkung: RNDr. Katarína Furmanová, Ph.D.

Wien, 25. Jänner 2023

[redacted]
Adam Böröndy

[redacted]
Renata Raidou

TU WIEN Informatics

Exploring and understanding the impact of machine learning choices on radiotherapy decision making

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Adam Böröndy, Bsc
Registration Number 01610133

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr. Renata Raidou
Assistance: RNDr. Katarína Furmanová, Ph.D.

Vienna, 25th January, 2023

[redacted]
Adam Böröndy

[redacted]
Renata Raidou

Erklärung zur Verfassung der Arbeit

Adam Böröndy, Bsc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 25. Jänner 2023

Adam Böröndy

Danksagung

Ich möchte mich bei meiner Betreuerin, Renata Raidou, für ihre Anleitung, Unterstützung und Geduld während des gesamten Verlaufs dieser Arbeit bedanken. Ihr Fachwissen und Feedback haben mir bei der Fertigstellung dieser Arbeit sehr geholfen. Ich möchte auch meine Dankbarkeit an Katarína Furmanová für ihre wertvollen Beiträge und Einsichten ausdrücken. Nicht zuletzt möchte ich mich bei meiner Familie und Freunden für ihre Unterstützung auf meinem akademischen Weg bedanken. Ihr Verständnis und ihre Ermutigung waren sehr geschätzt.

Acknowledgements

I would like to express my sincere appreciation to my supervisor, Renata Raidou, for her guidance, support, and patience throughout the course of this work. Her expertise and feedback helped me greatly in completing this thesis. I would also like to extend my gratitude to Katarína Furmanová for her valuable contributions and insights. Last but not least I would like to thank my family and friends for their support along my academic journey. Their understanding and encouragement have been greatly appreciated.

Kurzfassung

Bei der Planung der Strahlentherapie für Prostatakrebs ist eine genaue Beschreibung der Lage und Form der Beckenorgane ein entscheidender Faktor für die erfolgreiche Behandlung der Patienten. Die Behandlung erstreckt sich jedoch über einen längeren Zeitraum, in dem sich die Lage und Form der Organe erheblich verändern kann. Darüber hinaus ist das Ausmaß der Abweichungen von Person zu Person unterschiedlich. Aktuelle Publikationen haben dies untersucht, indem vorherige Patienten in Gruppen mit ähnlicher Organvariabilität unterteilt wurden. Die Verwendung dieser Erkenntnisse als Teil einer Vorhersage für die Organvariabilität bei neuen Patienten könnte die Behandlungsplanung verbessern und weiter personalisieren. Die statistischen und maschinellen Lernmethoden die in diesen Arbeiten eingesetzt werden, wurden bisher jedoch noch nicht gründlich und quantitativ ausgewertet und ihre Auswirkungen auf die abschließenden Vorhersagen wurden noch nicht genau untersucht. Diese Arbeit konzentriert sich auf eine bestimmte, von Furmanová et al. [FMCM+21] vorgeschlagene Implementierung dieser Ansätze und auf die quantitative Auswertung verschiedener Alternativen bei den verwendeten Methoden. Wir konzentrieren uns auf zwei Aspekte: die Auswirkungen der Verwendung verschiedener Methoden um den Form der Organe mathematisch zu beschreiben und die Auswirkungen von Änderungen bei den verwendeten Clustering-Algorithmen. Durch die Bereitstellung eines zusätzlichen Analyse Dashboards zur visuellen Bewertung der Auswirkungen der oben genannten Alternativen wollen wir eine mühelose und interaktive visuelle Interpretation der Auswirkungen der verschiedenen Änderungen ermöglichen. Dies soll den Entwicklern solcher Vorhersagealgorithmen dabei helfen, robustere Ansätze zu entwerfen. Als Fazit stellen wir fest, dass beim derzeitigen Stand der für die Analyse verwendeten Patientengruppe der Schwerpunkt auf der Auswahl geeigneter Methoden zur Beschreibung der Organformen liegen sollte, während die Auswirkungen der verschiedenen Clustering-Einstellungen auf die Vorhersage der extremsten Fälle von Varietät beschränkt sind.

Abstract

In prostate cancer radiotherapy planning, the accurate description of the position and shape of pelvic organs is a crucial part of successful patient treatment. However, the treatment is conducted throughout a long period of time, during which the position and shape of the organs might significantly vary. In addition, the amount of variation tends to differ for each individual. Recent visual analytics publications investigated this by partitioning past patients into clusters with similar variability. Using this as part of a prediction for the organ variability of new patients could improve and further personalize therapy planning. However, the statistical and machine learning methods employed in these works have not been thoroughly and quantitatively evaluated so far and their impact on the final predictions has not been assessed. This thesis focuses on taking a particular implementation of these approaches, proposed by Furmanová et al. [FMCM+21], and quantitatively evaluating the effects of using different alternatives for the employed methods. We focus on two aspects: the effect of using different shape descriptor methods and the impact of modifications in the clustering methods employed. By providing an additional visual analytics framework to visually assess the effect of the aforementioned alternatives, we aim to ensure an effortless and interactive visual interpretation of the impact of various modifications. This is anticipated to support the developers of said predictive algorithms in designing more robust approaches. As a result of our investigation we have highlighted potential issues and improved the initial implementation of the proposed workflow. We conclude that at the current stage of the patient cohort used for the analysis, the selection of appropriate shape description methods should be of main focus, while a notable impact of using different clustering methods is limited to the prediction of the most extreme cases of organ shape variations.

Contents

CHAPTER 1

Introduction

1.1 Problem Definition

1.1.1 Prostate Cancer and Treatment

Prostate cancer is the second most common cancer diagnosis and the fifth leading cause of cancer death in men worldwide according to the International Agency for Cancer Research (IARC) [IAR20]. In 2020, an estimated 1,414,249 new cases and 375,304 deaths related to prostate cancer have been recorded.

This type of cancer often appears as a slow-growing, low-risk cancer type and may be asymptomatic at the early stage with an indolent course, requiring only minimal interventions accompanied by active surveillance [Raw19]. Apart from these cases, there are three main treatment options commonly used for prostate cancer [KCM+13]:

- **Radical prostatectomy:** The surgical removal of the prostate and some of its surrounding tissue.
- **Prostate brachytherapy:** Also referred to as internal radiation therapy, which aims to kill cancer cells by temporarily placing radioactive sources directly in the prostate.
- **External beam radiation therapy:** Uses an external machine to deliver beams of radiation to the cancerous area. There are two types of beams commonly used to treat prostate cancer. Most radiation therapy devices use photon beams, which scatter a higher dose of radiation as they travel through the human body. Proton beams, on the other hand, have a lower scattering effect, which can reduce the radiation exposure to the non-target regions [RE19]. However, this type of treatment and the required devices are more expensive.

Which method is ultimately chosen depends on each individual case, but the decision is generally based on the main factors of age, cancer stage and risk class, with the patient also being involved in the decision-making process.

1.1.2 Toxicity

While all treatment options for prostate cancer are associated with specific risks and side effects, the focus of this work will be on external beam radiation therapy. When planning radiation therapy for prostate cancer, an accurate description of the location and shape of the pelvic organs is a critical factor for successful treatment of the patient. The use of suboptimal settings, which might overexpose healthy organs to radiation, could not only reduce the effectiveness of treatment but also negatively affect the healthy tissues [MLK+07, CMMH+17]. These negative side effects of radiation therapy are commonly referred to as *toxicity*. They range from temporary problems to long-term illnesses that severely affect patients' quality of life [MD14]. The acute effects of radiation, typically defined as occurring up to 6 months after treatment, include various urinary and bowel symptoms (e.g., frequency and urgency) and fatigue. Late toxicities include sexual dysfunction and persistent urinary and bowel problems, including intermittent rectal bleedings. An increased risk of secondary malignancies involving the irradiated areas (bladder, colon, and rectum) has also been reported after prostate radiation therapy [PP02, BCHR00]. As highlighted by Christie et al. [CSB15], potential toxicities following prostate cancer treatment are among the most common reasons patients regret their choice of treatment in retrospect.

To reduce the possibility and severity of toxicities, precise targeting of the cancerous area is crucial. To this end, the location of the pelvic organs is usually determined using Computed Tomography (CT) scans that capture their position and shape. However, CT scans provide only momentary images, while regular radiotherapy takes place over the course of several weeks. During this period, the position and shape of pelvic organs may change significantly from the planning phase [ZCM+99, ZLH+08, CMMH+17]. Furthermore, the extent of these variations tends to vary from patient to patient.

To account for organ variability, a safety margin is usually added to the organs so that possible variations in position and shape are accounted for [SRM+19]. However, this safety margin is generally a population-based estimate that does not take into account the patient's individual variability patterns and a more thorough analysis of this variability is required to assess the chance of toxicity [FMCM+21, RDCO+17].

1.1.3 Organ Variability Prediction using PREVIS

A series of recent publications have examined alternative approaches to analyzing and predicting the pelvic organ variability of individual patients [RCMA+18, GCMM+19, FGM+20, FMCM+21]. One of these new approaches (hereinafter: *PREVIS*), proposed by Furmanová et al. [FMCM+21], will serve as a basis for the research conducted in this thesis. The proposed approach and software implementation uses a set of cancer patients

from a previous cohort with known variability to generate personalized predictions for new patients. On a broad level, the approach consists of the following steps:

1. **Convert organ shapes captured by CT scans to shape descriptors**

   In order to work with organs, such as the bladder, rectum, and prostate, that have been captured by the CT scans, each organ must be represented in a way that adequately describes its shape. Using a mixture of abstraction methods, the organ segmentations in the CT scans are converted into mathematical shape descriptors that capture the presence of an organ at specific positions. Such an abstraction is also necessary for greater computational efficiency in later steps.

2. **Summarize shape variability for individual patients**

   For each patient in the past cohort, a series of registered CT scans from different stages of the treatment is available. In this step, the variability revealed by these scans is quantified. The variability of individual organs is measured as a deviation from the mean organ shape computed from all CT scans of the patient.

3. **Distinguish clusters and assign new patients**

   The most important part of the proposed approach is that it identifies clusters of patients with similar organ shapes and variability. Each new patient is then assigned to the most appropriate cluster, so that in subsequent steps the prediction is based only on information obtained from a selection of the most similar patients.

4. **Make predictions about patients with incomplete data**

   Once a new patient is assigned to the most similar cluster, patients from that cluster provide precedents and data for an individualized prediction of organ shape variability.

As this overview shows, the approach is based on a combination of several statistical and machine learning methods. However, the methods used have not yet been thoroughly and quantitatively evaluated. Several design decisions had to be made at each step, while viable alternative solutions were not tested. Therefore, to obtain a complete overview of the performance and potential of the proposed approach, a detailed analysis and evaluation of the workflow is required. The goal of this work is therefore to assess the impact of alternative choices in the workflow outlined above. The proposed changes focus on the two key steps: shape description (step 1) and clustering (step 2). Modifications to the shape description method could improve the precision with which the descriptors represent the original organ shapes. On the other hand, changes to the clustering settings could further improve matching similar patients and yield better predictions for new patients.

While some of the evaluated alternative solutions can be compared quantitatively by examining their effects on the predictive performance, many underlying changes remain

unclear when focusing only on this measure. Aspects such as changes in the cluster composition or precise changes in predicted shapes represent equally important information. However, these aspects are often qualitative in nature and require more detailed manual inspection to understand. To enable and support these use cases, a visual analytics application has been developed. This allows a detailed examination of various changes and their effects through visualizations and abstractions.

1.2 Research Questions

1.2.1 Main Research Question

The primary goal of this thesis is to evaluate alternative design options within the workflow proposed by Furmanová et al. [FMCM+21]. The main research question can be thus formulated as follows:

> *What is the impact of alternative choices, employed for the prediction of anatomical variability in PREVIS [FMCM+21], on the final outcome of the exploratory and predictive workflow?*

Corresponding to the workflow outlined in Section 1.1.3, there are two particular steps of interest. First, the shape descriptors used as input data for the analysis (step 1). Second, the employed clustering methods and their settings (step 2).

1.2.2 Input Data and Shape Descriptors

RQ 1 *What are the effects of modifications applied to the input data and shape description approach?*

The CT scans used as input data and the shape descriptors derived from them form the basis for the workflow. Their quality is therefore crucial for the feasibility of the entire approach. Therefore, the following sub-questions will be investigated:

RQ 1.1 *What are the effects of using different shape descriptors?*

The two most important criteria for the quality and usefulness of a shape description method in our case are, how well it allows a reconstruction of the original organ and how accurate predictions it enables. Therefore, the focus of this research question will be on these two main issues.

RQ 1.2 *What are the effects of introducing noise to the input data? How sensitive are various settings to inaccurate input data?*

While shape description methods provide a mathematical description of the organs captured by CT scans, these CT scans may be imprecisely labeled or

contain missing information. It is therefore of interest to see what impact these inaccuracies may have.

RQ 1.3 *What settings yield the best predictions for new patients with incomplete data? Does it change with increasing information available (i.e., further CT scans)?*

This research question addresses the boundary between shape descriptors and their subsequent use for clustering and prediction. With each additional CT scan for a patient, the information about organ variability also increases. With this increasing amount of information, the additional information obtained from other patients could become redundant. Therefore, the aim of this research question is to investigate how the prediction performance changes with an increasing number of CT scans and conduct initial assessments about the impact of using different clustering settings.

1.2.3 Clustering

RQ 2 *What are the effects of modifications applied to the clustering method and settings?*

The composition of the patient clusters has the largest effect on the predictions themselves. Any change in the clusters may lead to a change in the predictions. Therefore, it is of central interest to find optimal and reliable settings and to quantify the possible deviations in the prediction performance.

RQ 2.1 *What are the effects of using different parameterizations in the clustering (e.g., different similarity measures, different linkage methods)?*

The hierarchical clustering method applied in the original implementation of PREVIS relies on specific parameterizations that determine how the clusters are constructed. The purpose of this research question is to investigate the influence of different settings on the clusters and predictions.

RQ 2.2 *What are the effects of using a different clustering method (e.g., fuzzy or robust methods)?*

Besides hierarchical clustering, there are several other algorithms that could provide alternative clustering solutions. Therefore, it is of interest to see whether these provide different results with possibly improved prediction performance.

RQ 2.3 *How disruptive is the inclusion of a new observation with respect to existing clusters?*

Apart from the clustering method and settings chosen, the observations available in the dataset play the most important role. The inclusion or exclusion of certain

patients from the cohort could disrupt the previously identified clusters. This research question aims to investigate this effect and thus quantify the stability of the clusters.

1.3 Contribution

As described above, this work is directly based on a specific implementation by Furmanová et al. [FMCM+21]. By evaluating various aspects of this workflow, the results presented here are intended to demonstrate whether PREVIS can actually be used for the robust prediction of the anatomical variability of a patient within the course of radiotherapy treatment. It is aimed at assessing the choices made by the original implementation, and where possible improve it, thereby increasing its potential benefit for prostate cancer treatment.

Moreover, the developed visual analytics application will be applicable for the analysis of future enhancements of this or similar workflows. In particular, the visual analysis of cluster compositions under different parameterizations could prove to be useful for any project employing clustering methods, and support the developers of predictive algorithms to understand the impact of their choices on the outcomes of their models. In this way, our visual analytics approach can support them in designing more robust predictive solutions.

CHAPTER 2

Related Work

This chapter examines related research works, which have been conducted to answer similar research questions. These include directly related publications that served as a starting point for formulating our research questions. We will also pay special attention to examples of employing visual analytics in the context of these works. First we will focus on the domain of anatomical variability and provide a brief introduction to the problem setting. Then, we will shift our focus to the topic of shape descriptors—a commonly used method in this research area—and explore different approaches in this regard.

2.1 Anatomical Variability and Toxicity

2.1.1 Approaches in Radiotherapy

To obtain an accurate description of the location and shape of the pelvic organs, prior research studies have investigated various parts of the treatment process. As an example, the task is frequently approached from the viewpoint of the imaging techniques used as part of the treatment. Research studies indicate that opting for an MRI-guided treatment, as opposed to conventional CT scan-based ones, may be a cost-effective way to achieve better organ delineations and reduce toxicity [SLHC14, SDPHM20]. Others underscored the importance and necessity of adjusting the treatment settings as it progresses to account for anatomical variability [dCBP+18, GTF+11]. In particular, the use of image guidance prior to *each* treatment significantly reduced the risk to healthy tissues. Finally, an important role is played by the safety margins applied around the pelvic organs. An illustration displaying the concept of safety margins is shown in Figure 2.1, originally presented by Schlachter et al. [SRM+19]. In this figure, different boundaries represent the following volume concepts:

- Gross Tumor Volume (GTV): The extent of the tumor visible by means of imaging techniques.

- Clinical Target Volume (CTV): Extends the GTV by encompassing additional microscopic extensions into healthy tissues.
- Internal Target Volume (ITV): Takes also the uncertainties due to organ motion into account.
- Planning Target Volume (PTV): Takes both the uncertainties due to organ motion as well as uncertainties due to setup error into account, thus applying safety margins around the CTV.
- Organs at Risk (OAR): Healthy tissues possibly affected by the treatment.
- Treated Volume (TV): The volume intended to receive at least a minimal level of radiation dose appropriate for the purpose of the treatment.
- Planning Organ at Risk Volume (PRV): The OAR with additional safety margins, similarly to the relation of CTV and PTV.

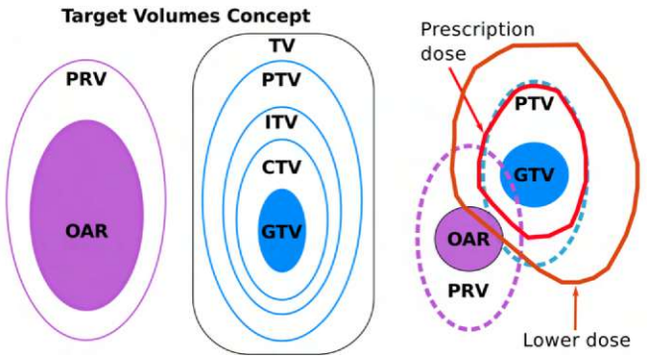


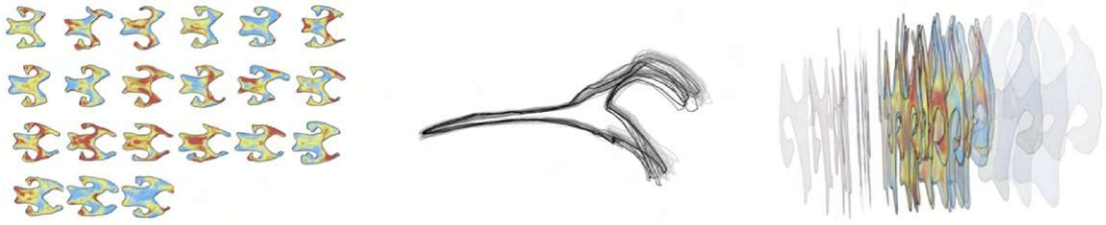
Figure 2.1: Volume concepts used in radiation therapy [SRM+19]

In a review of commonly used guidelines and settings, Yartsev and Baumann [YB16] noted that the selection of safety margins is not consistent across facilities and drew the conclusion that optimal protocols for quality assurance procedures are needed.

2.1.2 Visual Analytics

As the above examples show, the expected variations in shape and position of the pelvic organs play an important role during prostate cancer radiotherapy treatment. To analyze and visualize the variability of organs and shapes in general, a number of research studies have proposed possible frameworks. Busking et al. [BBP10] was one of the first to develop an interactive visual analytics application for exploring shape variations. For visualizing multiple shapes in a single view and thus highlighting their variability, they have proposed three different approaches. A comparison of these can be seen in Figure 2.2. In Figure 2.2 (a) three-dimensional shapes are visualized with their variation around

a mean shape highlighted by colors. In Figure 2.2 (b) the variability is captured by comparing only specific contours of the observations. In Figure 2.2 (c) a combination of these approaches can be seen, where all contours are stacked along one dimension.



(a) Side-by-side (b) Overlaid contours (c) Shape stack

Figure 2.2: Shape evolution views proposed by Busking et al. [BBP10] for general shape variability analysis

Other publications focused specifically on the relationship between organ shape variability and segmentation errors yielded by different algorithms [RBGR18, VLBK+13]. Klemm et al. [KLR+13] focused on human spines and developed a tool to visually examine different spine shapes and search for clusters of patients with similarly shaped spines. In a later publication [KOJL+14], this approach was extended to include more details about the patients, such as their age and gender, to enable an analysis across these dimensions as well. The implemented visualization dashboard can be seen in Figure 2.3.

Figure 2.3: Visualization dashboard by Klemm et al. [KOJL+14] for the analysis of spine shapes

To describe the variability of pelvic organs specifically, recent research studies proposed a series of approaches. Raidou et al. [RCMA+18] (hereinafter: *Bladder Runner*) focused exclusively on the bladder and developed a visual analytics tool to explore the shape variations of individual patients throughout the treatment period. It also allowed an evaluation of the impact of these variations with respect to the accuracy of the delivered dose during radiotherapy treatment. In this approach, coverage confidence levels derived from the variations are visualized rather than the actual observations (see Figure 2.4).

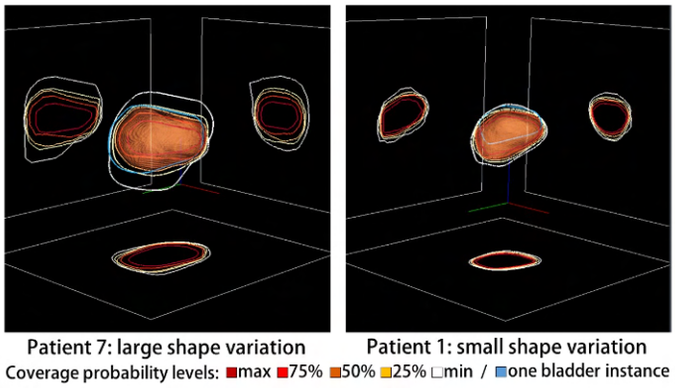


Figure 2.4: Individual level bladder shape variability visualization using probability coverage levels in Bladder Runner [RCMA+18]

For a cohort-wide comparison of organ variations, an abstracted view using bubble glyphs has been proposed (see Figure 2.5). In this case, each row represents the observations belonging to a single patient, with the area of the bubbles representing the bladder volume. Furthermore, the authors employed clustering techniques to distinguish patients with similar bladder shape variability. These groupings are highlighted by color coding individual observations.

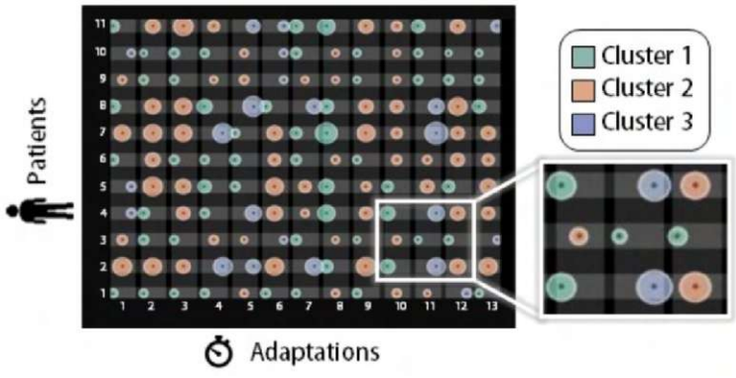


Figure 2.5: Cohort-wide bladder shape variability and cluster comparison in Bladder Runner [RCMA+18]

On the one hand, the approach shown in Figure 2.4 enables a more detailed analysis of organ variations in individual patients. On the other hand, the implementation shown in Figure 2.5 provides a means for cohort-wide analysis, but is limited in terms of highlighting detailed variations of individual patients or clusters of patients. A similar approach has been proposed by Grossmann et al. [GCMM+19] (see Figure 2.6), which extended Bladder Runner to include other pelvic organs. In this case, observations are again color-coded according to their cluster assignment, with cluster-specific variability patterns visualized in separate views.

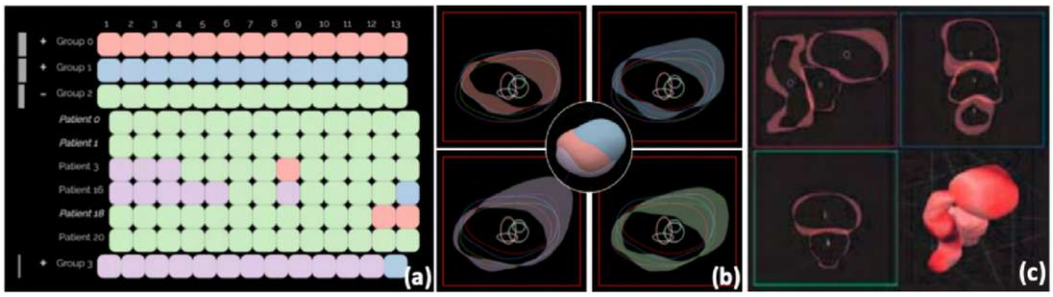


Figure 2.6: Pelvic organ shape variability and cluster comparison in Pelvis Runner [GCMM+19]

Building mainly on these two approaches, Furmanová et al. [FGM+20] (hereinafter: *VAPOR*) included a more detailed analysis of the correlation between shape variations and possible toxicities. In particular, they extended the visualization of pelvic anatomy variability over the treatment period by also incorporating the delivered dose into the analysis. Finally, in a subsequent work by Furmanová et al. [FMCM+21], the prediction workflow evaluated in this thesis has been proposed. A detailed specification of this workflow was presented in Section 1.1.3. In terms of the employed visual analytics solution, in this implementation, patients were grouped based on their overall pelvic organ shape and shape variability patterns. The resulting groupings are indicated by color coding on the left side of the dashboard, next to the patient identifiers (see Figrue 2.7). In addition, the difference of each time step compared to the first—planning CT scan (pCT)—is visualized by a heatmap of gray-scale colors. On the right side of the dashboard, the patients are displayed in a RadViz plot [HGP99], with the distribution of the points giving a first indication about the similarity between them. The same color coding as on the left side of the dashboard highlights the groupings identified by the clustering of the patients. Based on these, the cluster-specific patterns are visualized by a thumbnail indicating the dominant shape patterns in each cluster.

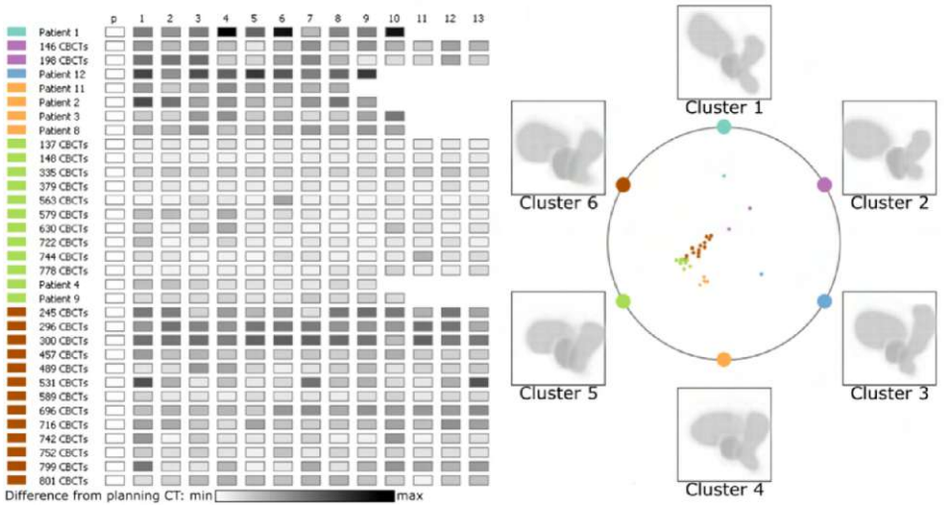


Figure 2.7: Pelvic organ shape variability and cluster comparison in PREVIS [FMCM+21]

The above publications are related in many ways, with most of them offering alternative approaches and extensions to address similar tasks. PREVIS, the most recent work in this series, offers a prototype that combines most of the advantages of earlier works. It provides an implementation that is applicable to multiple pelvic organs, employs clustering techniques to find patients with similar patterns, and at the same time provides a way to make predictions about the expected organ shape variations.

2.2 Shape Descriptors

In research areas that deal with three-dimensional shapes, shape description methods are often employed to produce abstracted representations for the actual shapes. Focusing on the topic of pelvic organs, we can see that the publications mentioned in Section 2.1 use varying methods. Bladder Runner relied on a 14-dimensional descriptor [PI+97], with the elements of the descriptor capturing various features of the shapes, including information about their principal axes, volume, compactness etc. However, this method proved to be too simplistic for more complex shapes, especially for non-spherical ones. Therefore, later extensions of the approach in Pelvis Runner and VAPOR employed linearization strategies using scanline and hilbert curves [Hil35], followed by Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) as dimensionality reduction techniques. Reiter et al. [RBGR18] used a spherical harmonic-based descriptor [KFR03] as shape description method for the pelvic organs. In an additional step, they also applied the same dimensionality reduction techniques on the descriptors as in Pelvis Runner and VAPOR. The use of such techniques allowed the generation of two dimensional descriptions for the organs and a subsequent visualization of their similarity using scatter plots. This in turn, facilitated use cases such as outlier detection. Finally, the approach presented in PREVIS relied on a probabilistic shape

description method put forward in a work by Akgül et al. [ASYS06]. In contrast to other methods, this representation calculated a probability for the presence of an organ at specific target points. One of the advantages of this approach was that the location of target points were uniform for all patients, which enabled a direct comparison between them. Furthermore, it allows for an easy inclusion of new patients to the cohort without the need to recalculate the descriptors for all patients.

2.3 Summary

Although all the above described approaches make specific, informed choices, e.g., for shape descriptors and for clustering methods to use, none of them explicitly investigates all possible alternatives, nor provides a thorough assessment thereof. Yet, if approaches such as PREVIS are to be integrated into clinical decision-making processes, a thorough analysis of the entire choice space is required. Furthermore these works frequently present diverging results, demonstrating that a thorough evaluation is necessary for each application scenario.

CHAPTER 3

Methodology and Approach

This chapter presents the methodology used to investigate the research questions posed in Section 1.2, divided into the two topics of statistical testing and evaluation, and visual analytics. With regards to the statistical part, the most important aspects of the workflow used to evaluate and improve the results of previous research will be explained. With respect to visual analytics, the key design decisions for the interactive dashboard that has been developed as part of this work will be summarized. In particular, we will highlight which research questions were addressed on each dashboard page and which tools have been employed to do so.

3.1 Testing and Evaluation

The first part of this thesis focuses on the testing and evaluation of the workflow proposed by Furmanová et al. [FMCM+21]. For this, the following steps were taken:

1. **Scope Definition:** While the existing implementation is a prototype for the proposed approach, it was not designed with the flexibility to support all the evaluation scenarios proposed in this work. Moreover, only certain parts of the implementation were required for our investigations. After identifying the necessary parts required to answer the research questions defined in Section 1.2, these were moved outside of the prototype and assessed in a more efficient environment. This allowed an optimization of the workflow and ensured reasonable run-times in subsequent simulations. Furthermore, it also served as an evaluation of the reproducibility of the results presented by Furmanová et al. [FMCM+21] and highlighted potential problems.

2. **Adjustment of the workflow:** Based on the knowledge gained during the first step, specific parts of the workflow were adapted or improved. This included

both performance-related optimizations but also enhancements to the core of the approach. Additional modification made the workflow more flexible to support the use of different settings according to the research questions presented in Section 1.2.

3. **Testing and evaluation:** Finally, the evaluation of the prediction workflow took place. This was done using a leave-one-out cross-validation approach, by simulating each patient once as a new patient with incomplete data, similarly to the evaluation process used by Furmanová et al. [FMCM+21]. This stage of the evaluation focused on quantitative aspects, such as the overlap of the predicted and the target output shapes. While valuable insights have already been gained from this evaluation, the purpose of this step was also to provide a pre-computed dataset for the exploration using the visual analytics. This also allowed the visual analytics part of the approach to run smoothly without having to re-run the simulations each time.

To ensure that modifications within the workflow aligned with the original purpose of the software, these steps were conducted in close collaboration with the developers of the prototype.

3.2 Visual Analytics

In the second part of this thesis, a visual analytics workflow has been developed to support the investigation and interpretation of the results provided by the first part. To follow the general workflow of going from a broader overview to individual patients, a visualization dashboard with three different pages has been implemented. Below, a short description about each individual page highlights its purpose. Additional illustrations highlight the implemented visualizations and their interactions on each page. These also include an overview of the steps required for each visualization, defined similarly to the approach proposed by Brehmer et al. [BM13]. We also highlight the connection of each dashboard page to the research questions defined in Section 1.2. In this context, it should be noted that research questions **RQ 1.1** and **RQ 1.2** are specific in that they depend fundamentally on the input data used for the application. Therefore, e.g., the effects of different shape descriptors on the clustering can be explored by using different input datasets.

1st Page—Cohort Overview: This page serves as a starting point for the visual exploration of the patient cohort. It provides insights into how certain settings influence the division of the cohort into separate clusters. Its main goal is to help understand the differences between the various clustering settings and to select the optimal ones. While a two-dimensional scatterplot gives an initial indication of the distribution and similarity of patients in the cohort (Figure 3.1, Patient Similarity), a Sankey diagram is used to visualize the cluster hierarchy as a function of the number of clusters (Figure 3.1, Cluster Composition). An additional diagnostic line chart facilitates the evaluation of the quality of different clustering solutions (Figure 3.1, Cluster Validity). These tasks are directly

related to the research question **RQ 2**: *What are the effects of modifications applied to the clustering method and settings?* and all of its sub-questions defined in Section 1.2.

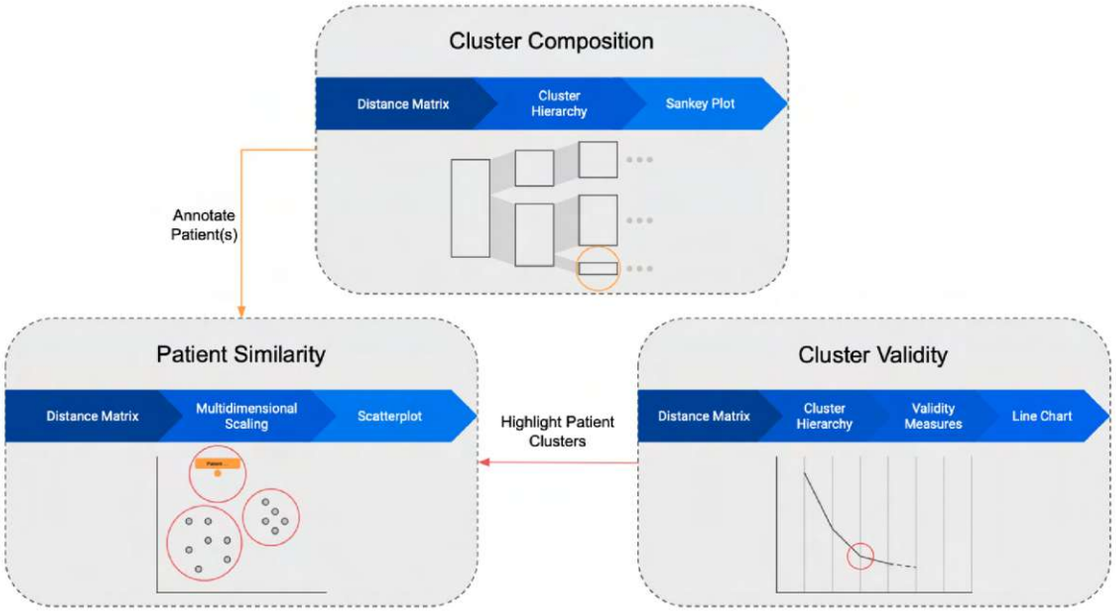


Figure 3.1: 1st Page–Cohort Overview (supporting an exploration of clustering settings and addressing research questions **RQ 2.1**—**2.3**)

2nd Page—Patient/Cluster Overview: In a second step, the focus shifts to individual patients of interest and their relationship to the other patients in the cohort. In particular, it is of interest, which patients in the cohort are grouped together with the patient of interest under different clustering settings. In addition, this step also provides information about the predictive performance of each setting with respect to the patient of interest, which is based on the pre-calculated evaluation dataset from the first part of this work. The exploration is mainly supported by a parallel coordinates plot [ID09] that can be restricted to any combination of variables, with the order of adjacent variables highlighting their correlations (Figure 3.2, Settings and Performance Overview). Thus, this page focuses on the impact of various clustering settings with respect to a single patient of interest and supports the investigation of research question **RQ 1.3**: *What settings yield the best predictions for new patients with incomplete data? Does it change with increasing information available (i.e., further CT scans)?*.

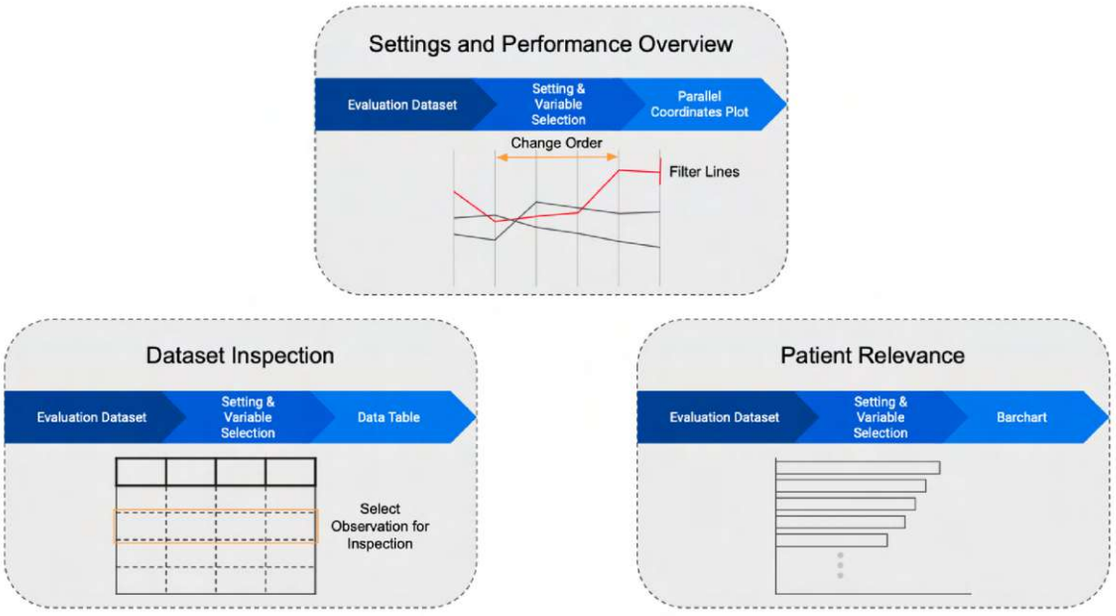


Figure 3.2: 2nd Page–Patient/Cluster Overview (supporting an exploration and comparison of the prediction performance and addressing parts of research question **RQ 1.3**)

3rd Page—Prediction Inspection: Finally, once a single patient of interest and a specific clustering setting has been selected, the third page enables an inspection of the precise predictions. This includes quantitative aspects such as the similarity (Figure 3.3, Cluster Overview) and overlap (Figure 3.3, Shape Overlap) between the organs of the patient of interest and individual cluster patients. In addition, this page also allows for a qualitative assessment of the overlap between the target and predicted shapes by visually inspecting the underlying three-dimensional shapes (Figure 3.3, Shape Comparison). An additional performance chart provides key insights into the relation of the number of available CT scans and the predictive performance (Figure 3.3, Performance Chart), addressing further aspects of research question **RQ 1.3**.

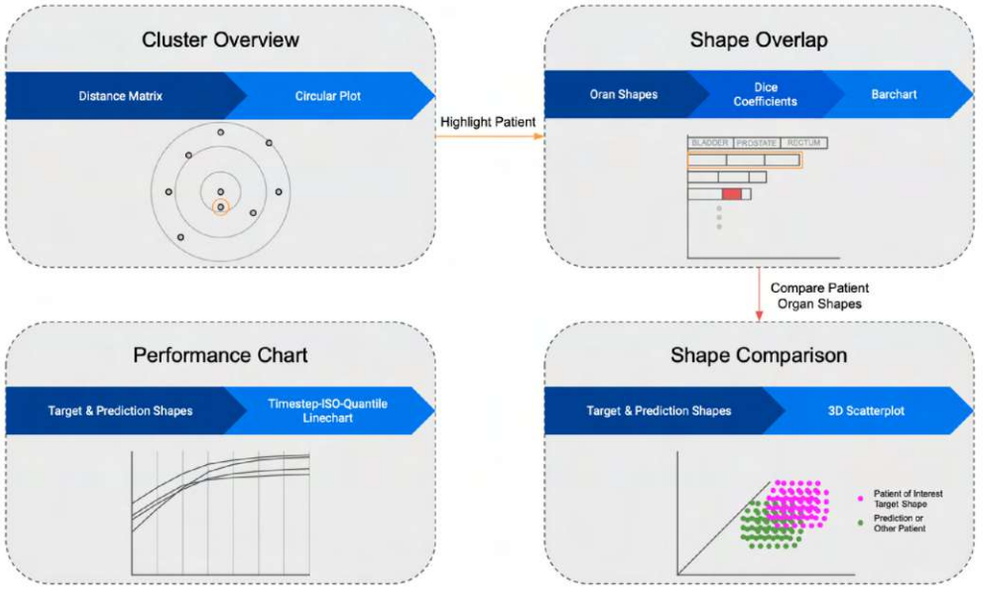


Figure 3.3: 3rd Page—Prediction Inspection (supporting an exploration of individual predictions and addressing parts of research question **RQ 1.3**)

CHAPTER 4

Implementation

4.1 Software Implementation

4.1.1 Shape Descriptors

Motivation and Requirements

The CT scans and their derived organ delineations provide the foundation for the workflow presented in this work. However, in order to describe the variability of the organs in a mathematical way, a more quantitative representation is required. Therefore, shape description methods are applied to generate one-dimensional vectors per patient, organ and timestep, capturing the size and shape of the organs. In our use case, an optimal shape description method should fulfill for the following four requirements:

1. The descriptors should be directly *comparable* for their similarity. This means that the measured features and the dimensions of the shape descriptors should be identical across different observations.

2. There should be a way to *revert* the descriptors and accurately *reconstruct* the input shapes. Furthermore, using the same reconstruction workflow with identical settings for different observations, should yield reconstructions with comparable quality. This consistency is critical in order to formulate reliable guidelines for a cohort wide analysis of the observations.

3. The descriptor generation workflow should be easily *scalable*, such that new patients can be added to the cohort without requiring an update of all other descriptors.

4. Finally, the quality of the shape descriptors should be *controllable*. Increasing their complexity should yield higher quality descriptors that capture more and more details about the input shapes.

Descriptor Generation

Input Data The data used as input for the shape descriptors consists of the *source points* provided by the organ delineations. To demonstrate this, Figure 4.1 visualizes the bladder from the planning CT scan of patient *137 CBCTs*. Figure 4.1 (a) visualizes all source points in three dimensions, while Figure 4.1 (b) shows only a single slice of the scan. It is worth noting that the patient cohort used for the evaluations in later sections was obtained from two different medical institutions, each of which provided observations with different resolutions. The source points are distributed uniformly across a grid, with the distance between them determining the resolution of the shapes. Thus, the source points create a voxel space in three dimensions, with the points being the center of the voxels.

(a) 3D shape (b) 2D sample slice

Figure 4.1: Sample bladder source points

Prior Research As presented in Section 2.2, related research studies have implemented several solutions to generate shape descriptors for the input data presented above. While these methods were optimal for their specific use case, none of them provides a solution, which would fulfill all requirements defined in this section. The 14D shape descriptors used in Bladder Runner [RCMA+18] were suitable for the bladder and potentially other spherical-like shapes, but not sophisticated enough for more complex, non-spherical shapes (as opposed to our use case in general). Pelvis Runner [GCMM+19] and Vapor [FGM+20] employed linearization techniques to describe the organs in an abstracted way. Both implementations applied principal component analysis to create descriptors of lower dimensions. However, this approach requires a dimension reduction step to be rerun for the whole cohort with each new patient (as opposed to requirement 3). Apart from efficiency considerations, this could also lead to changing descriptors across all patients in the cohort. Furthermore, none of the above mentioned approaches implemented an approach for a reconstruction of the organ shapes from the shape descriptors (as opposed to requirement 2). An alternative approach, addressing all requirements for the shape descriptors was presented in PREVIS [FMCM+21]. In this case a set of specific *target points* were used as abstraction points for the input shapes. These target points were

identical for all observations, while the presence of an organ in their surrounding was captured by assigning specific probabilities to them.

Bounding Box and Target Points To illustrate the concept of target points, this step reduces it to a simplified, two-dimensional input shape. However, all the steps presented here, can be applied in the same way to any three-dimensional shape. Consider the shape in Figure 4.2 as the source points of a theoretical input shape.

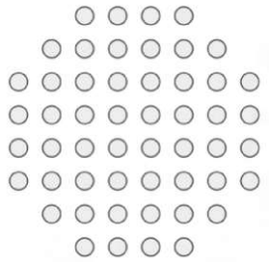
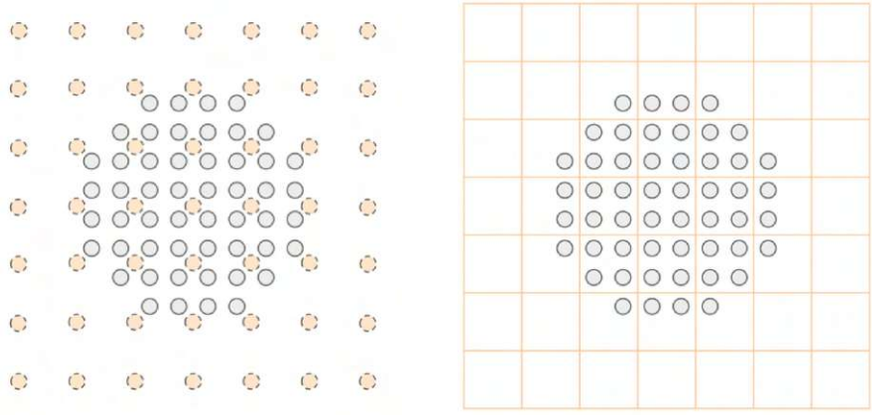


Figure 4.2: Sample shape source points

The first step in the construction of the shape descriptor is to calculate a bounding box that covers the entire shape. This bounding box determines the dimensions of the space, in which we want to describe an organ. Next, a set of target points needs to be distributed in the form of a regular grid within this space. The distance between the target points defines the resolution of the resulting shape descriptor. In case of the original implementation of PREVIS this was set to 15mm. If we visualize the surrounding neighborhood of these target points, they form a grid structure. Figure 4.3 (a) illustrates the distribution of target points, while Figure 4.3 (b) their corresponding space coverage as a grid.



(a) Target points (b) Target grid

Figure 4.3: Regular grid of target points

It is worth noting that the dimensions of the bounding box are not patient-specific, but are based on the entire patient cohort. This means, that the bounding box must be large enough to cover the organs of *all* patients within the cohort. As a general rule, this also leads to an increase in the number of target points required, as the organs of each individual patient from the cohort vary in size and location. At the same time, it is expected that the size of the required bounding box stabilizes with an increasing cohort size, as outliers in terms of organ shape, size and position become less frequent. However, this approach ensures that the shape descriptors of different patients rely on the same target points and have the same dimensions, which makes them directly comparable for their similarity. As a next step, a value needs to be assigned to each of the target points, representing the probability for the presence or absence of an organ in that region.

PREVIS Shape Descriptors To assign specific probabilities to the target points, the original implementation of PREVIS relied on an approach proposed by Akgül et al. [ASYS06]. This method is based on kernel density estimation, in which an occupancy probability is assigned to each of the previously defined target points, representing the probability of the target points being part of the organ. This is done using a kernel function that estimates probabilities depending on the frequency of organ source points in the area surrounding each target point. After tailoring it to the use case, PREVIS used a Gaussian kernel with the density estimate shown in Equation 4.1.

$$f_S(t_n \mid O) = (2\pi)^{-m/2} \sum_{k=1}^{K} |H_k|^{-1} \exp\left(-\frac{1}{2}(t - s_k)^T H_k^{-2} (t - s_k)\right) \tag{4.1}$$

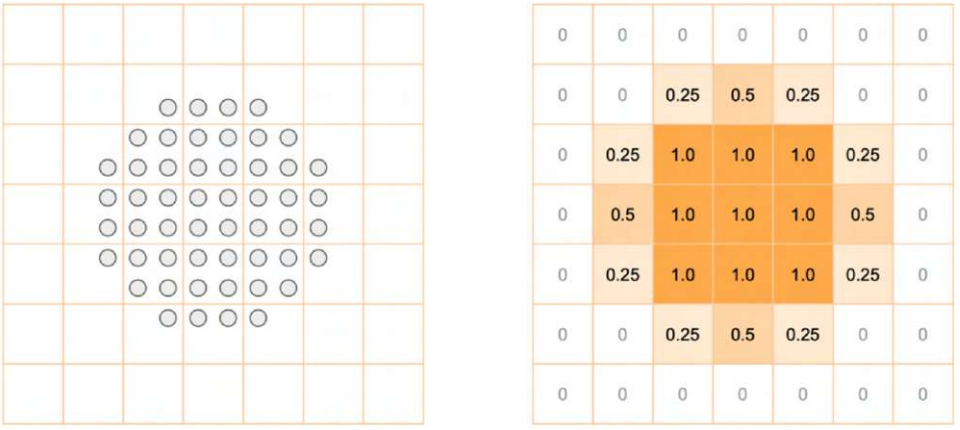
Here, $\{t_n \in \mathbb{R}^m\}_{n=1}^N$ represents each individual target point determined according to the approach described previously. Moreover, $\{s_k \in \mathbb{R}^m\}_{k=1}^K$ consists of the source points that make up the organ delineations. Note that the only parameter that requires an estimation is the bandwidth parameter $H \in \mathbb{R}^{m \times m}$, which serves as a smoothing parameter. For the computation of the bandwith, PREVIS relied on an estimate given by Scotts's rule [HMSW04]. With the assumption that the bandwidth parameter remains the same for the entire shape, Equation 4.1 can be rewritten as follows:

$$f_S(t_n \mid O) = C \sum_{k=1}^{K} \exp\left(-\frac{1}{2}(t - s_k)^T H^{-2} (t - s_k)\right) \tag{4.2}$$

With $C = (2\pi)^{-m/2}|H|^{-1}$ being a constant value. With this simplification, it becomes visible that the estimate for each target point depends on its squared distance to the organ source points, adjusted by the bandwidth parameter.

Resolution Based Shape Descriptors An alternative solution for assigning probabilities to the target points is to essentially reduce the resolution of the input shape. Since the target points are arranged on a regular grid, they already provide a perfect

coverage of the shape space. Thus, the shape space can be divided into non-overlapping regions according to the regular grid defined by the target points. Next, the individual probabilities can be estimated based on the density of source points in the neighborhood of each target point. The highest density corresponds to the highest probability of 1, with all other target points scaled accordingly. For demonstration purposes, we show a two-dimensional simplification of our approach in Figure 4.4. In Figure 4.4 (a) the regular grid defined by the target points is overlayed on the source points of the input shape. In Figure 4.4 (b) the respective coverage probabilities of the target points are visualized.



(a) Target grid (b) Descriptor probabilities

Figure 4.4: Sample shape descriptor

In a final step the target points can be flattened in to a one-dimensional vector, which is practical for use cases, such as calculating the similarity of different shape descriptors. By flattening the matrix of values in Figure 4.4 (b), one would get the following vector of probabilities representing the final shape descriptor for the input shape in Figure 4.2:

$$d = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0.5 & 0.25 & 0 & \cdots & 0 \end{bmatrix}^T$$

Shape Reconstruction

In the context of the workflow presented in this work, an important aspect of the shape description method used is its ability to accurately reconstruct the original input shape. In our approach, an up-sampling of the shape descriptor vector is used for this step. This is facilitated by the property of the shape descriptor, that it abstracts the input shape into a regular grid of probability values. We illustrate the principles of this process on the resolution based shape descriptor derived from the two-dimensional sample input shape presented in Figure 4.2. In a first step, the two-dimensional grid described by the descriptor vector gets reconstructed, which is shown in Figure 4.4 (b). Then, the

up-sampling process continues by inserting additional rows and columns into the grid structure, thereby doubling its resolution. This is illustrated in Figure 4.5.

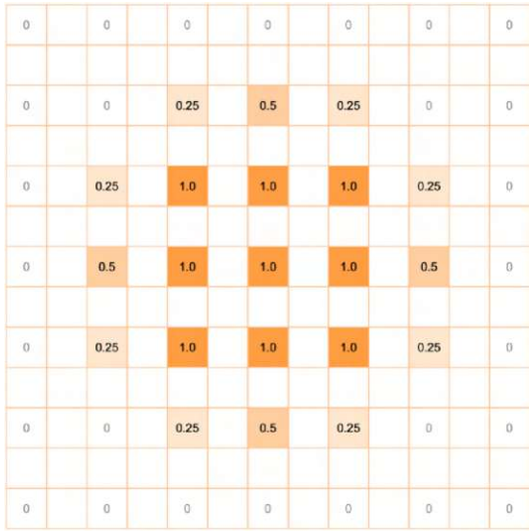


Figure 4.5: Up-sampled dimensions

Next, a linear interpolation of the probabilities is conducted, during which the newly inserted cells are imputed with an average of the surrounding probabilities, thus ensuring a smooth transition of the values. Depending on the desired outcome, this up-sampling procedure can be repeated several times, with each step increasing the resolution of the shape. Figure 4.6 shows the resulting shape after two up-sampling iterations, which was the setting used by PREVIS and in later evaluations presented in this work as well.

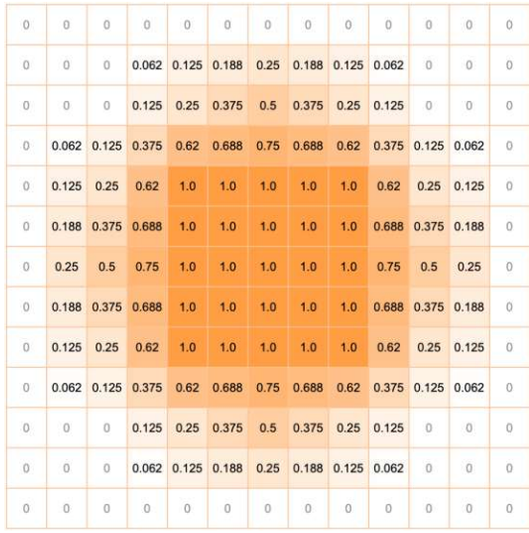


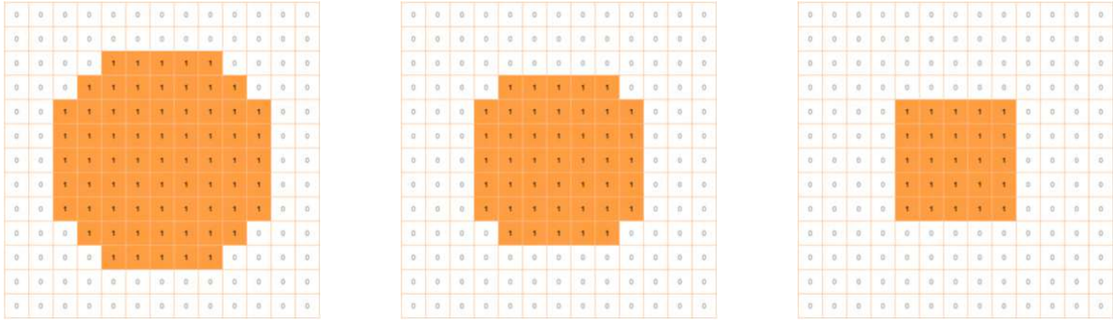
Figure 4.6: Up-sampled probabilities

Finally, a smoothing of the up-sampled shape is conducted. During this step, each point within the space is updated as the mean of its direct neighborhood. The final result is shown in Figure 4.7. It is worth noting that in extreme cases where the upsampled shape does not have a large enough core with probabilities of 100%, the smoothing process could decrease all values below 1. However, this would generally indicate that the shape descriptor was suboptimally dimensioned in the first place, and has only a few target points fully covered by the organ. To avoid this phenomenon, the smoothing procedure can be extended with a division of the resulting values by the maximum probability, ensuring that the probabilities range from 0 to 1.

0	0	0.007	0.021	0.042	0.062	0.069	0.062	0.042	0.021	0.007	0	0
0	0	0.021	0.062	0.125	0.188	0.208	0.188	0.125	0.062	0.021	0	0
0.007	0.021	0.083	0.188	0.312	0.417	0.444	0.417	0.312	0.188	0.083	0.021	0.007
0.021	0.062	0.188	0.375	0.562	0.688	0.708	0.688	0.562	0.375	0.188	0.062	0.021
0.042	0.125	0.312	0.562	0.778	0.896	0.903	0.896	0.778	0.562	0.312	0.125	0.042
0.062	0.188	0.417	0.688	0.896	1.0	1.0	1.0	0.896	0.688	0.417	0.188	0.062
0.069	0.208	0.444	0.708	0.903	1.0	1.0	1.0	0.903	0.708	0.444	0.208	0.069
0.062	0.188	0.417	0.688	0.896	1.0	1.0	1.0	0.896	0.688	0.417	0.188	0.062
0.042	0.125	0.312	0.562	0.778	0.896	0.903	0.896	0.778	0.562	0.312	0.125	0.042
0.021	0.062	0.188	0.375	0.562	0.688	0.708	0.688	0.562	0.375	0.188	0.062	0.021
0.007	0.021	0.083	0.188	0.312	0.417	0.444	0.417	0.312	0.188	0.083	0.021	0.007
0	0	0.021	0.062	0.125	0.188	0.208	0.188	0.125	0.062	0.021	0	0
0	0	0.007	0.021	0.042	0.062	0.069	0.062	0.042	0.021	0.007	0	0

Figure 4.7: Probabilities after smoothing

To finalize the shape reconstruction, a cut-off value is required, above which we consider a target point to be a part of the output shape. Different cut-off values (also referred to as *ISO* values) and their corresponding shape reconstructions can be seen in Figure 4.8.



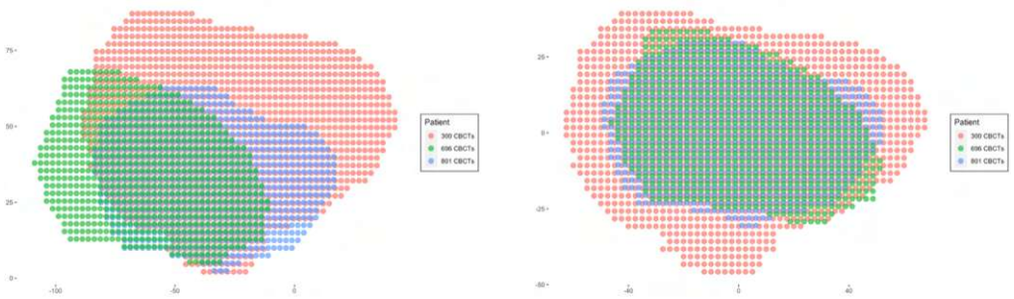
(a) ISO 0.25 (b) ISO 0.50 (c) ISO 0.75

Figure 4.8: Reconstructed shapes for different cut-off (ISO) values applied on the probability map depicted in Figure 4.7

Centered Shape Descriptors

To ensure that CT scans from different patients and time steps comparable, each CT scan is centered with respect to the prostate. However, no additional adjustment is made for other organs, such as the bladder and rectum. Therefore, these organs are analyzed in terms of their relative position to the prostate. While this approach is unproblematic for the analysis of a single patient, it could pose difficulties for comparisons between different patients, as the organ placement might not be comparable. It is worth noting, that one argument for this approach is, that the relative position of the organs could affect their variability. For example, the proximity and contact of the bladder to the prostate could limit its space for shape variations—or, on the contrary, be influenced by the variations of the surrounding organs.

Nonetheless, further improvements might be gained by centering each organ of the individual patients according to the first, treatment planning CT scan. This way, each organ could be analyzed independently from the position of other organs. A comparison of the two approaches, using a slice of the bladder CT scan from three radiotherapy patients, can be seen in Figure 4.9. Comparing the prostate centered setting in Figure 4.9 (a) with the individually centered setting in Figure 4.9 (b), it is visible, how the additional centering of the shapes facilitates a direct comparison between them. Furthermore, since the organs are located in a more compact space in Figure 4.9 (b), the dimension of the bounding box and the number of target points required for the shape descriptor generation would decrease simultaneously as well.



(a) Prostate centered (b) Individually centered

Figure 4.9: Organ centering comparison

4.1.2 Patient Descriptors

The patient cohort used for the analysis presented in this thesis consists of 33 patients. For each of these patients 9–14 CT timesteps provide temporal acquisitions of their organ shape and position. Figure 4.10 highlights the changes throughout these timesteps for a sample patient's bladder, by comparing the probabilities for a single slice of their shape descriptor.

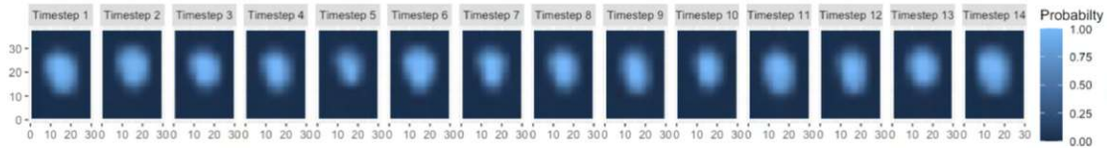
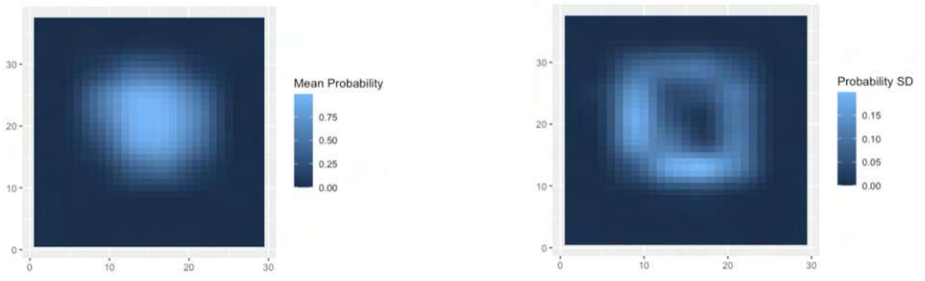


Figure 4.10: Bladder shape descriptor slice comparison for patient *137 CBCTs*

To describe the overall patterns of each patient, an aggregation of their available shape descriptors is performed. This aggregation consist of two key components. First, the mean probability at each target point for each individual organ captures its average shape and position. Second, the standard deviation at each specific target point captures the observed variability of the shapes. The resulting components for the patient bladder presented in Figure 4.10 are visualized in Figure 4.11. Here, Figure 4.11 (a) presents the mean shape of the organ, while Figure 4.11 (b) highlights the variations of the organ.



(a) Mean Shape (b) Standard Deviation

Figure 4.11: Patient descriptor for patient *137 CBCTs* (bladder components)

Finally, the mean and standard deviation descriptor for the three organs of bladder, rectum and prostate are concatenated into one single *patient descriptor*. Repeating the above steps for each patient in the cohort yields a set of patient descriptors, which we next compare for their similarity and split into clusters.

4.1.3 Clustering

In the context of this work, various approaches for the clustering of the patient descriptors are assessed and compared in terms of their suitability for the proposed prediction workflow. The description in this section provides a theoretical overview of the underlying principles behind each of these approaches.

Hierarchical Clustering

We first introduce the agglomerative hierarchical clustering algorithm [LLSE11], which was also the method of choice for the original implementation of PREVIS. In this approach, each observation is initialized as its own cluster, resulting in n clusters, with n being

the number of patients in our case. Next, two of these clusters are merged into a new one, reducing the overall number of clusters to $n-1$. This step is iteratively repeated until each observation is merged into a single large cluster. During this workflow, the selection of the two clusters for each merger is determined by two key parameters. First, a distance method determines a way to quantify the distance between two observations—in our case the patient descriptors. Second, a linkage method is needed to determine how to measure the similarity between two distinct clusters. In particular, it is of interest which observations within a cluster are considered when comparing the distance between clusters. Based on these two parameters, the two most similar clusters are identified and selected for a merger. In the scope of this work, the following distance methods are evaluated:

- *Manhattan distance*: The sum of absolute distances between the Cartesian coordinates of two vectors.

$$D(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad \text{where } x_i \text{ and } y_i \text{ denote identical positions in the vectors } x \text{ and } y \text{ of length } n$$

- *Euclidean distance*: The square-root of the sum of squared distances between the Cartesian coordinates of two vectors.

$$D(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \text{where } x_i \text{ and } y_i \text{ denote identical positions in the vectors } x \text{ and } y \text{ of length } n$$

- *Minkowski distance*: An generalization of the Manhattan and Euclidean distances, where the power p must be manually given. Here, $p = 1$ is equivalent to the formula of the Manhattan distance, while $p = 2$ to the formula of the Euclidean distance.

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad \text{where } x_i \text{ and } y_i \text{ denote identical positions in the vectors } x \text{ and } y \text{ of length } n \text{ and } p \text{ being a freely choosable exponent}$$

- *Binary distance*: Vectors are regarded as binary bits, with non-zero elements handled as a bit of 1 and zero elements as 0. After this transformation, a distance is calculated as the proportion of positions where only one of the vectors takes a bit value of 1.

$$D(x,y) = \frac{1}{n} \sum_{i=1}^{n} x_i \neq y_i \quad \text{where } x_i \text{ and } y_i \text{ denote identical positions in the vectors } x \text{ and } y \text{ of length } n \text{ after transformation into binary bits}$$

- *Canberra distance*: A weighted version of the Manhattan distance, meaning that it is more sensitive to changes if both coordinates are near to zero.

$$D(x,y) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{|\mathrm{x}_i - \mathrm{y}_i|}{\mathrm{x}_i + \mathrm{y}_i} \text{ where } x_i \text{ and } y_i \text{ denote identical positions in the vectors } x \text{ and } y \text{ of length } n$$

- *Chebyshev/Maxmimum distance*: A measure, in which the distance is given by the largest difference along any coordinate dimension.

$$D(x,y) = \max_i (|x_i - y_i|) \text{ where } x_i \text{ and } y_i \text{ denote identical positions in the vectors } x \text{ and } y \text{ of length } n$$

In addition, the following linkage methods are evaluated in conjunction with the distance methods to generate alternative clustering results:

- *Single linkage*: Calculates a similarity value between clusters A and B based on the two closest observations in them.

$$D(A,B) = \min_{a \in A, b \in B} \{d(a,b)\} \text{ where } a \text{ and } b \text{ are elements of } A \text{ and } B \text{ respectively}$$

- *Complete linkage*: Calculates a similarity value between clusters A and B based on the two most distant observations in them.

$$D(A,B) = \max_{a \in A, b \in B} \{d(a,b)\} \text{ where } a \text{ and } b \text{ are elements of } A \text{ and } B \text{ respectively}$$

- *Average linkage*: Calculates a similarity value between clusters A and B based on the average distance of all pairs of observations between the cluster.

$$D(A,B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b) \text{ where } a \text{ and } b \text{ are elements of } A \text{ and } B \text{ respectively, with } |A| \text{ and } |B| \text{ denoting the number of elements in each of them}$$

- *McQuitty linkage*: Calculates a similarity value between clusters A and B based on the average distance of all pairs of observations in the union of the two clusters.

$$D(A,B) = \frac{1}{(|A|+|B|)(|A|+|B|-1)} \sum_{x,y \in A \cup B} d(x,y) \text{ where } |A| \text{ and } |B| \text{ denote the number of elements in them, and } x \text{ and } y \text{ being elements of their union}$$

- *Centroid linkage*: Calculates a similarity value between clusters A and B based on the distance between the mean vectors of the two clusters.

$$D(A, B) = d(\bar{a}, \bar{b}) \text{ where } \bar{a} \text{ and } \bar{b} \text{ denote the mean vector of } A \text{ and } B$$

- *Median linkage*: Calculates a similarity value between clusters A and B based on the distance between the median vectors of the two clusters.

$$D(A, B) = d(\tilde{a}, \tilde{b}) \text{ where } \tilde{a} \text{ and } \tilde{b} \text{ denote the median vector of } A \text{ and } B$$

- *Ward's minimum variance linkage*: Calculates a similarity value between clusters A and B based on the increase in variance when merging them. This variance measure amounts to a weighted squared distance problem between the cluster centers.

$$D(A, B) = \frac{d(\bar{a}, \bar{b})^2}{1/|A| + 1/|B|} \text{ where } \bar{a} \text{ and } \bar{b} \text{ denote the mean vector of } A \text{ and } B \text{ and } |A| \text{ and } |B| \text{ the number of elements in them}$$

The progression of the algorithm described above is often visualized by a dendrogram, highlighting the cluster mergers as hierarchy. An example dendrogram for the patient cohort used in this work in combination with the application of Euclidean distance and complete linkage can be seen in Figure 4.12. Furthermore, this hierarchy provides us with a clustering solution for all possible number of cluster—ranging from 1 to n. In Figure 4.12 we have highlighted how the division of the cohort into 3 clusters would look like.

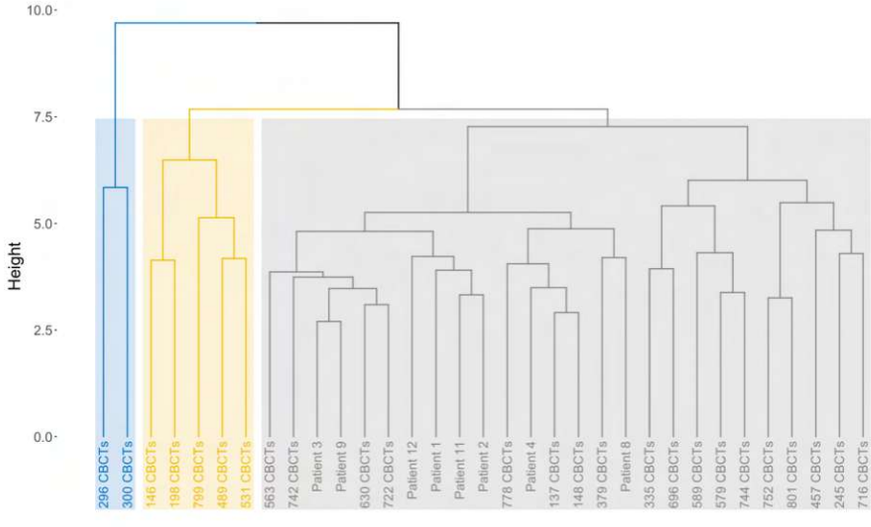


Figure 4.12: Sample patient cluster dendrogram (using Euclidean distance and complete linkage)

k-means Clustering

In addition to the hierarchical clustering algorithm, this work extends the evaluation to other alternative clustering approaches as well. One such widely used alternative method is the k-means clustering algorithm [HW79]. A major difference to hierarchical clustering is that this method can be used with the patient descriptors directly as input. In addition, this approach requires the number of clusters as a predefined parameter k for initializing the algorithm. Once the input data and the number of clusters are determined, the algorithm follows the following iterative approach to optimize the cluster assignment of the observations:

1. Initialize the position of the k cluster centers. This is generally done by selecting random observations from the dataset.
2. Assign each observation to the nearest cluster center.
3. Update the position of the cluster centers as the mean of the observations assigned to the cluster.
4. Repeat Step 2 and Step 3 until the position of the cluster centers and the cluster assignment of each observation does not change.

A common problem with this algorithm is that the returned solution may depend on the position of the cluster centers that were first initialized [PLL99]. Therefore, certain solutions may not correspond to the overall best cluster assignments. A solution of this type is often referred to as a local optimum. To avoid such cases, the algorithm can be repeated several times and the most frequent result can be considered as the overall, global optimum. Considering the iterative optimization task described above, it can be expressed as a mathematical minimization task as well. In a global optimum solution, the cluster centers are positioned such that the sum of distances between the cluster centers and the observations in the clusters is minimized. This measure, called the within-cluster scatter, is reduced with each iteration of the algorithm until convergence. Using the Euclidean distance as a distance measure between observations, we can therefore define the objective function in Equation 4.3, which is minimized by the algorithm described above:

$$\arg\min_{k} \sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_{i\cdot} - \overline{\boldsymbol{x}}_k\|^2 \text{ where } \overline{\boldsymbol{x}}_k \text{ is the mean vector of cluster } k \tag{4.3}$$

k-medoids Clustering

A frequently used variation of k-means is the k-medoids algorithm. While this approach still uses the same input data and also requires the number of clusters as a predefined parameter, it has two major differences to k-means. First, this approach always chooses

one of the observations as the center for a cluster. Second, instead of minimizing the sum of squared Euclidean distances, it minimizes the sum of pairwise dissimilarities, i.e., the sum of dissimilarities between the observations in the cluster and their centroid. This peculiarity of k-medoids that it does not square the distances involved in the minimization task, makes it more robust to noise and possible outliers. In the context of our work, we will rely on the widely used Partitioning Around Medoids (PAM) algorithm, described by Kaufman et el. [KR90b]. As a measure of dissimilarity, we again use the Euclidean distance to achieve better compatibility settings between k-means and k-medoids. The cost function to be minimized in this case is defined by Equation 4.4:

$$\arg\min_{k} \sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_{i.} - \boldsymbol{m}_k\| \text{ where } \boldsymbol{m}_k \text{ is the medoid of of cluster } k \tag{4.4}$$

Fuzzy c-means Clustering

While the above mentioned models provide categorical, or so called hard clustering results, there exist methods that assign individual observations to multiple clusters at the same time. In this case the membership of each observation to a cluster is expressed as coefficients that are generally designed in a way to add up to 1 overall (i.e., they can be interpreted as percentage wise memberships). One such clustering method is fuzzy c-means [BEF84], which directly builds on the principles of k-means but incorporates the above mentioned features. Its working mechanism is best highlighted by the underlying objective function, shown in Equation 4.5:

$$\arg\min_{k} \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} \|\boldsymbol{x}_{i.} - \boldsymbol{c}_k\|^2 \tag{4.5}$$

Here, u_{ik}^{m} is the cluster dependent weight of individual observations, while $\boldsymbol{c}_k$ represents the weighted cluster center as described by Equation 4.6. Furthermore the exponent $m \in (1, \infty)$ is a variable that determines the degree of fuzziness. While increasing the exponent will result in fuzzier clusters, in the limit $m \to 1$ the memberships will converge to either 0 or 1:

$$m_{kj} = \frac{\sum_{i=1}^{n} u_{ik}^{m} x_{ij}}{\sum_{i=1}^{n} u_{ik}^{m}} \tag{4.6}$$

Model Based Clustering

As an additional approach, we also evaluate the use of model-based clustering [FR02] for our use case. In general, this method assumes that the dataset was generated using a combination of multiple statistical models, with each component having different parameterizations. The statistical model of choice is typically a multivariate normal distribution, with the assumption that the different components—being essentially the

cluster structures we want to identify—differ in terms of location and covariance structure. These unknown parameters are then estimated by an expectation maximization algorithm. Although this is not explicitly the setting in our workflow, the underlying methods could still provide new insights. In addition, because the dimension of the shape descriptors are very large and computing a covariance matrix would require estimating many parameters, we impose constraints on the covariance structures. Such constraints are often required and are also included in the software implementations of the algorithm. These constraints primarily address two aspects of the clusters: their shape (e.g., being restricted to a spherical shape) and their size (e.g., being equal across clusters), both controlled by the covariance structure involved in the estimation.

4.1.4 Generative Model

The core of the prediction workflow presented in this work is the generative model, which is used to capture the predominant patterns of organ variation in a set of input patients. This extracted information is then applied to generate a large set of possible organ shape variations for the new patient.

The clustering methods described in the previous section introduced different approaches to identify a group of the most similar patients for a new patient. In this step, we use the information about this subset of patients—as well as the limited information already available for the new patient—as input to the generative model. First we calculate the deviation from the mean shape of each patient for each CT scan. This results in a set of individual organ variations measured by a change in the probability of the shape descriptors. A collection of such variations is shown in Figure 4.13. This plot summarizes some of the typical shape variations that can be observed for the organ of the bladder, restricting the visualization to a single slice of the organ. In *Timestep 4* we can observe an increase in the volume of the organ compared to the mean shape, while in *Timestep 14* a shrinkage of the organ. Finally, in *Timestep 5* a shift in the position of the organ resulted in a decrease of the probabilities on one side and an increase on the other side of the organ.

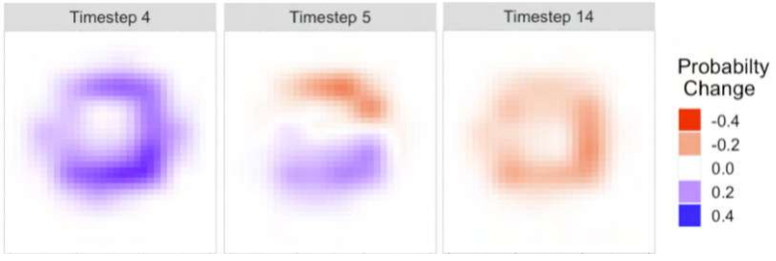


Figure 4.13: Sample variations of a bladder around the mean shape (patient *148 CBCTs*)

The first part of the generative model summarizes the variability patterns of all the selected patients and CT scans in a mathematical way. To do this, we rely on an approach proposed by Budiarto et al. [BKS+11]. In this method, principal component analysis

is applied to extract the main modes of variation from a set of shape descriptors. As previously described, we subtract the patient mean shape from each individual shape descriptor and thus end up with centered data, which we will denote by d_c. These vectors can be used directly as input to calculate the covariance matrix C as a means to summarize the shape variations in the selected patient group, as shown in Equation 4.7.:

$$\mathcal{C} = \frac{1}{N-1} \sum_{i=1}^{N} d_c \cdot d_c^t \tag{4.7}$$

Using this covariance matrix, we next conduct an eigenvalue decomposition under the constraint that the returned eigenvectors should be normalized to unit length. To then generate new samples of organ shape variations, we use normally distributed data as scalars (c) to multiply the eigenvectors (q) with. The mean of the normal distribution is 0, while its variation is determined by the square root of the respective eigenvalues. The approach can be restricted to the first l largest eigenvalues, which capture the desired amount of variation in the data, as shown in Equation 4.8:

$$d_{cnew} = \sum_{l=1}^{L} c_l q_l \tag{4.8}$$

To illustrate the output of this approach, Figure 4.14 presents a set of sample shape variations generated by the workflow introduced above.

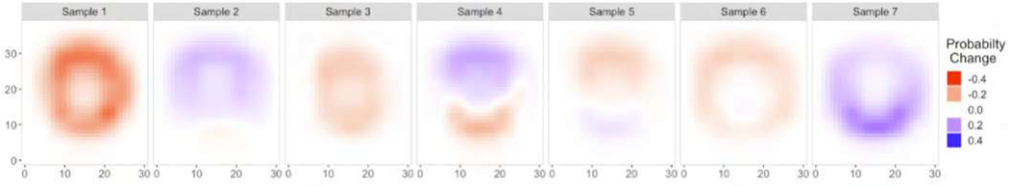


Figure 4.14: Sample bladder variations generated by the generative model

4.1.5 Making Predictions

Once a sufficiently large set of suitable shape variations has been generated (by default 1000 samples), they are aggregated and used for the final shape prediction for the new patient. To describe our prediction approach we first introduce the desired target shapes and then explain how the prediction workflow attempts to match them.

Target Shapes

The target shapes, which are later used as the ground truth for evaluating the predictions, describe the shape of the organs under specific conditions. While the primary aspect for assigning appropriate safety margins to an organ is its size and position under the

most extreme conditions (i.e., that it is covered even at its largest volume), we extend our prediction workflow to different cases of organ variation as well. Our goal is to provide a workflow that can predict both the average shape and the expected shrinkage or enlargement of an organ. To account for all of these cases, we categorize different types of variation by computing specific quantiles of variation in the probabilities of the shape descriptors at individual target points. To illustrate this, Figure 4.15 shows the same section of the bladder descriptors at different time steps for patient *137 CBCTs*. Each target point in these visualizations represents a probability for the presence of the bladder in that region. Next, the probabilities at individual target points are aggregated by calculating certain quantiles of their values. Figure 4.16 shows the resulting shapes for three specific quantile settings.

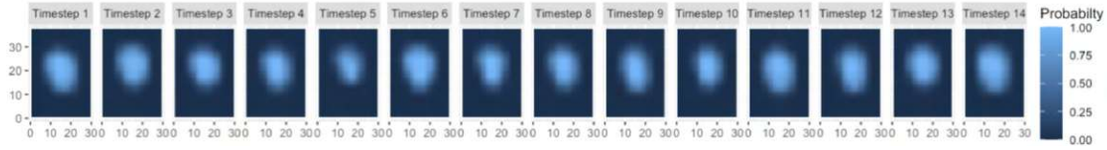


Figure 4.15: Bladder shape descriptor slice comparison for patient *137 CBCTs*

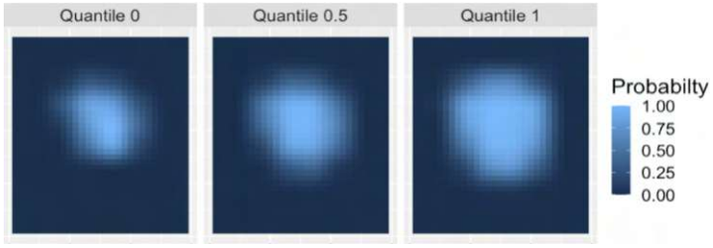


Figure 4.16: Bladder shape variation quantile sample for patient *137 CBCTs*

Predicted Shapes

As for the prediction part of our workflow, the input data for this step consists of the mean shape of the new patient and the samples of shape variations provided by the generative model described in the previous section. As part of the prediction workflow we first aggregate the shape variations according to a selected quantile setting. The variations themselves are measured as a change in the probabilities relative to the mean probability at individual target points. Assuming that the probabilities increase or decreases to roughly the same amount around the mean a quantile of 0.5 is equal to 0 across all target points. Above this level, higher quantiles describe an increase in organ volume, with a quantile of 1 describing the largest probability increase for all target points. Accordingly, quantiles below 0.5 represent a shrinkage relative to the mean shape. Figure 4.17 illustrates this on some real world examples using our workflow.

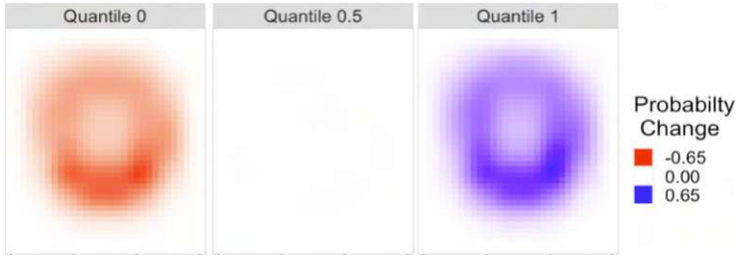


Figure 4.17: Predicted bladder shape variation quantiles for patient *137 CBCTs*

These shape variations can be interpreted as the expected shape deformations of the new patient compared to its mean shape. The final shape predictions can thus be obtained by adding the generated shape variation quantiles to the mean shape of the new patient. To provide a visual overview of this approach, Figure 4.18 summarizes our workflow and shows the three edge cases of greatest, least, and no variation added to the mean shape of the new patient *137 CBCTs*. While these predictions must be interpreted as probabilities, a cutoff value can be chosen to get a categorical prediction with an actual organ shape. Such categorical predictions are also used for the assessment of the prediction performance by calculating a Dice coefficient [Sor48, Dic45] between the predicted and the target shape for individual quantile settings.

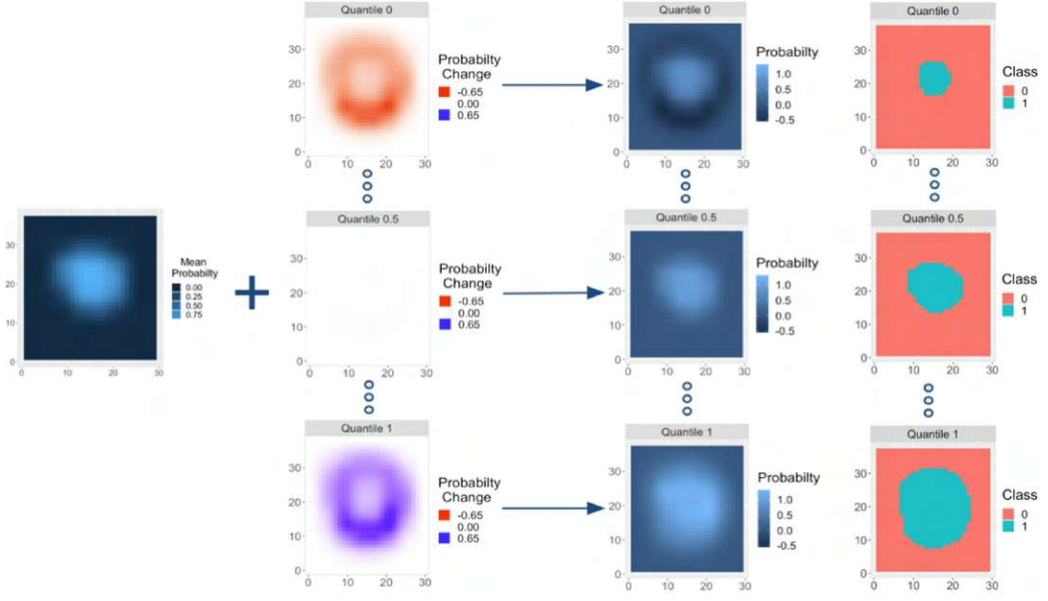


Figure 4.18: Shape prediction workflow example for patient *137 CBCTs*

It should be noted that one disadvantage of this method is that more complex patterns of variation, in particular shifts in organ position, are not well represented by individual quantiles. A decrease on one side of the organ is included in a quantile below 0.5, while the increase on the other side is included in a separate quantile above 0.5. Therefore,

the interpretation of individual quantiles is primarily limited to conclusions about the enlargement or shrinkage of the volume.

4.2 Evaluation Workflow

In order to answer the research questions defined in Section 1.2, our approach includes a detailed statistical evaluation, with the goal of gaining quantitative insights. As part of this, we frequently compare or rank the methods investigated. In this section we outline the approach used for these evaluations on the level of individual research questions, with the actual findings presented in Section 5.

4.2.1 Shape Descriptors - Research Question 1.1

RQ 1.1: *What are the effects of using different shape descriptors?*

This thesis compares and evaluates the two types of shape description methods described in Section 4.1.1. First, the approach used in the original implementation of PREVIS. For these shape descriptors, a pre-calculated dataset is utilized, which was also used for the evaluations presented by Furmanová et al. [FMCM+21]. Second, the resolution-based descriptor proposed in this work. These are calculated according to the approach presented in Section 4.1.1. With respect to these two alternative shape description methods, their ability to reconstruct the original input shape is a crucial aspect of their quality. One way to measure the overlap between the input and reconstructed shapes is by calculating their respective Dice coefficient [Sor48, Dic45]. This measure quantifies the overlapping area of the input (X) and reconstructed (Y) shapes as a percentage of the overall area covered by them, as shown in Equation 4.9:

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{4.9}$$

To evaluate and compare the shape reconstruction capability of the two approaches, both types of descriptors are up-sampled to the same resolution as the original input shapes. Furthermore, the up-sampling is finalized with various ISO values as possible cut-off settings. Finally, for each up-sampled shape, the Dice coefficient as a measure of overlap between the input and reconstructed shape is calculated. This process yields a set of possible reconstructions, on the basis of which we can evaluate the two key parts of this research questions. First, it is of interest whether one of the shape description methods provides a superior reconstruction quality. For this task, the optimal ISO values—leading to the highest dice coefficients—are compared across the two shape description methods. Second, it is of interest whether the optimal ISO values remain consistent across patients and time steps. This is crucial because later evaluations are based on the assumption that specific ISO settings provide a comparable shape reconstruction quality overall.

4.2.2 Shape Descriptors - Research Question 1.2

RQ 1.2: *What are the effects of introducing noise to the input data? How sensitive are various settings to inaccurate input data?*

While the shape descriptors derived from the CT scans are used for essentially all of the evaluations presented in our work, their quality does not only depend on the appropriate choice of the shape description method but clearly on the quality of the underlying input data as well. Therefore, we must take into account various issues that the input CT scans could have as well. An example for such an aspect is the possibility for systematic differences in settings used for the acquisitions. For example to address the differences in the positioning of the patients, all of the CT scans are centered around the organ of the prostate. Similarly, differences in the resolution of the CT scans is handled by the abstraction into shape descriptors of identical dimensions. Apart from these general quality aspects, certain issues could affect the quality of the shape descriptors on the level of individual patients. Imprecise organ delineations could lead to inaccurate organ models. However, the quality of the delineations is hard to quantify with possible human errors hard to simulate in an evaluation setting. In this research question we focus on a related issue, frequently affecting individual CT scans, namely missing slices in the organ annotations. These can happen due to several reasons, including human errors during the annotation of the individual slices or the transmission of the data. Since missing slices are interpreted as empty space within the organ, the derived shape descriptors could lack in quality as a result. Therefore, we investigate this issue in two stages. First we randomly exclude individual slices from the CT scan and then inspect to what degree the derived shape descriptors are still capable of providing a representation for the actual, full organ. In a second stage we increase the amount of missing information by iteratively excluding an increasing number of neighboring slices from the CT scans and repeating the same analysis. To align this simulated setting with possible real world cases, we restrict our analysis to missing slices along the plane of acquisition in the CT scans.

4.2.3 Shape Descriptors - Research Question 1.3

RQ 1.3: *What settings yield the best predictions for new patients with incomplete data? Does it change with increasing information available (i.e., further CT scans)?*

This research question is addressing the boundary between the two main components of interest, namely the shape descriptors and the clustering methods employed. As highlighted in Section 4.1.2, the shape descriptors and the number of CT scans available for new patients have a direct impact on the patient descriptors that capture the organ position and shape variability patterns of each patient. Subsequently, changes in the patient descriptors may lead to changes in the clustering results, which in turn may affect the predictions. Therefore, in this research question we are primarily interested in *how* the prediction performance changes as the number of available CT scans increases. Furthermore we are also interested in seeing, to what degree the prediction performance depends on using different clustering settings. To answer these questions, a leave-one-out

cross-validation process is performed, simulating individual patients as incoming patients with incomplete data. As part of this process, an increasing number of CT scan time steps are included to generate predictions and then compared for their differences. Finally, the process is repeated with different clustering settings to investigate their effects on the prediction performance. The results of this workflow were then stored in an evaluation dataset for further inspection. This dataset was not only used for direct comparisons and statistical evaluations between clustering settings, but is also an important input for the visualization interface presented in the next chapter. It is worth noting that the generation of this evaluation dataset is computationally the most intensive part of the statistical evaluation presented in this thesis. Therefore, parallelization techniques were heavily utilized in the implementation of the required workflow, relying mainly of the *R* package extensions of *doParallel* and *foreach*. In addition, to ensure reproducibility of results, parts of the workflow involving randomly generated data (e.g., the normally distributed scalars generated during the generative model described in Section 4.1.4) were controlled by predefined random seeds.

4.2.4 Clustering - Research Question 2.1/2.2

RQ 2.1: *What are the effects of using different parameterizations in the clustering (e.g., different similarity measures, different linkage methods)?*
RQ 2.2: *What are the effects of using a different clustering method (e.g., fuzzy or robust methods)?*

As mentioned in Section 4.1.3, the original implementation of PREVIS relied on agglomerative hierarchical clustering to divide the patient cohort into subgroups. However this algorithm has the limitation that the results depend on the choice of parameters, which can significantly affect the composition of the clusters. Therefore, in **RQ 2.1**, we evaluate the use of alternative distance and linkage methods for hierarchical clustering. To do this, we iteratively modify these settings and evaluate the changes in cluster compositions as well as the changes in the predictive performance of the workflow. Of key interest here are aspects such as the extent to which certain settings can identify outliers in the patient cohort or whether the size of the clusters is approximately equal. As an extension, in **RQ 2.2** we evaluate alternative clustering methods using the same evaluation process.

4.2.5 Clustering - Research Question 2.3

RQ 2.3: *How disruptive is the inclusion of a new observation with respect to existing clusters?*

While the focus of this work is the evaluation of the prediction workflow first presented by Furmanová et al. [FMCM+21], it builds on a series of related publications which proposed a clustering of prostate cancer patients according to their organ shape similarity. The clustering methods compared in **RQ 2.1** and **RQ 2.2** all share the common goal of identifying groups of similar patients, and accomplish it using different approaches, leading to slightly different results. Many clustering methods however are known to

be sensitive to changes in the data, where the inclusion or exclusion of even single observations can disrupt the previously identified groupings. Moreover, such disruptions in the cluster assignments could further propagate into the prediction workflow and lead to changes in the prediction performance for new patients as well. This research question simulates this issue, by iteratively excluding individual observations and inspecting their effects on the cluster assignments. It is of interest to see how large of a disruption individual patients cause under different clustering settings—measured for example by the number of patients that switched clusters as a result of the modifications.

4.3 Visual Analytics Application

To facilitate the interactive and flexible investigation of the research questions presented in Section 1.2, a visual analytics interface has been developed. As proposed in Section 3.2, this application divides the analytical workflow into three separate pages, each with a unique purpose. This section presents the specifics of the implementation and also discusses possible alternative solutions that were considered but discarded.

4.3.1 Technical Implementation

Along with the R-based implementation of the evaluation workflow, the visual analytics interface was also developed in R. More specifically, the visual analytics application presented here is an R-Shiny[1] application. This package extension was explicitly designed to allow the development of standalone web applications with R code running in the backend. For our use case, this allows the proposed statistical analysis to be executed in R, while the R-Shiny extension handles the information exchange for presenting the results in a web application. There are two key concepts to be aware of in order to understand the technical implementation of the analytical dashboard presented in this section. First, the structure of the application is divided into two separate components. A user interface object determines the structure and layout of the front end of the web application. This typically includes methods to define input variables, such as selecting values from a predefined set, or setting new values through slider inputs, etc. In addition, output elements can be defined as placeholders for plots and tables whose contents are defined or generated later. The second component of the application is the server side, which defines the logic and backend for the application. While in our use case most of the statistical analysis is presaved in an evaluation dataset, the server side takes over the task of filtering the relevant data based on the input variables and converting it into the desired visualizations. To demonstrate this component structure, an example application is presented here. Starting with Code Block 4.1, in the first step we load the required package extension and define the user interface component of the application. The *sidebarLayout()* function defines the layout of our application, which is divided into a sidebar panel and a main panel that can be accessed individually by the corresponding functions. Within the sidebar panel, we define a single input, which we give the identifier

[1] https://shiny.rstudio.com

n. For this input, we select a slider as a tool to set a number between 1 and 1000, with the default value set to 100. In the main window, we define a single output called histogram, which we will later construct as a graph.

```
library(shiny)

ui <- fluidPage(
    sidebarLayout(
        sidebarPanel(
          sliderInput(inputId = "n",
                      label = "Number of values:",
                      min = 1,
                      max = 1000,
                      value = 100)
          ),
        mainPanel(
          plotOutput(outputId = "histogram")
        )
    )
)
```

Code Block 4.1: Example user interface component of Shiny application

Next, in Code Block 4.2, on the server side of the application, we establish a link between the input variable and the desired output. Therefore, we now define the output as a histogram that visualizes a randomly selected set of normally distributed values. The key aspect here is that the number of values drawn from this distribution depends on the input n, which is set manually by the user of the application. This property is the second important feature for the application presented in this paper. Each time an input variable is changed by the user, it triggers an update of the elements depending on its value. In a final step, we use the user interface and server components and run the application. The resulting web application of this introductory example is shown in Figure 4.19, with the layout shown in a minimized view of the browser for illustration purposes.

```
server <- function(input, output) {

    output$histogram <- renderPlot({
        valueSample <- rnorm(n = input$n,
                             mean = 0,
                             sd = 1)
        hist(x = valueSample,
             main = "Histogram of the generated values",
             xlab = "",
             ylab = "")
```

```
11        })
12  }
13
14  shinyApp(ui = ui, server = server)
```

Code Block 4.2: Example server component of shiny application

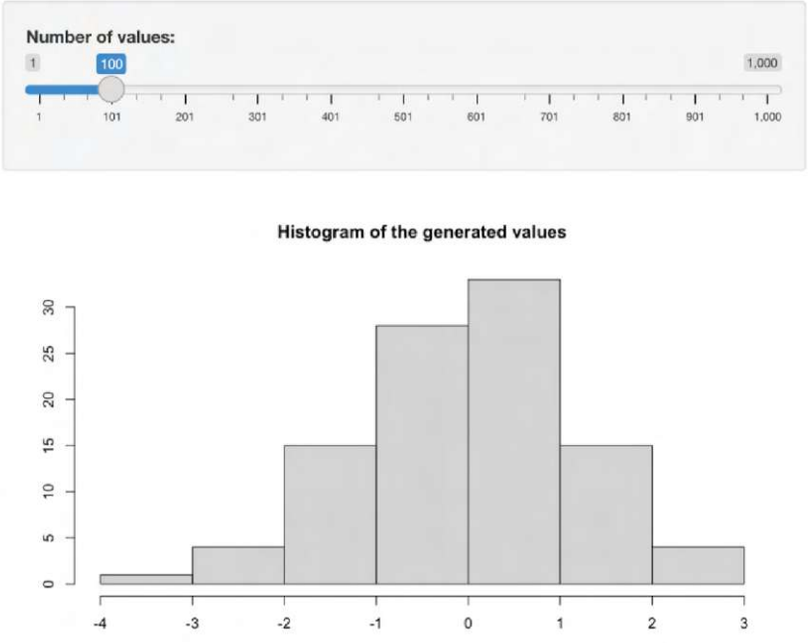


Figure 4.19: Introductory Shiny application example

While certain additional enhancements allow further customization of the application's layout, for example to limit reactivity between dependent elements in the application, the structure presented above serves as the foundation for the application presented in the following sections. To gain detailed insight into the exact implementation of the application, the source code has been made publicly available in a GitHub repository[2].

4.3.2 1st Page—Cohort Overview

The first page, shown in Figure 4.20, provides an initial cohort-wide overview of the patient cohort. This is also the landing page after launching the application and is intended as the starting point for the analytical workflow that was considered when designing the dashboard. Figure 4.20 shows a screenshot of this application page, with the numbers indicating the various elements incorporated.

[2]https://github.com/adamborondy/masterthesis

Element (1) This element shows a scatter plot of all patients in the cohort. When designing this visualization, the main goal was to represent the cohort of patients in an abstracted way that also depicts the similarity between them. This is important because we aim to identify clusters of patients, and thus similarity is an essential feature for us. To meet these requirements, we relied on the patient descriptors, which are constructed in such a way that the vectors can be directly compared for their similarity using different distance methods (see Section 4.1.2). For this visualization, we relied on the Euclidean distance as the distance measure and computed a distance matrix that captures the pairwise distance between all patients in the cohort. In a next step, we used multidimensional scaling to obtain a two-dimensional projection of this information. Alternative dimension reduction methods, such as t-SNE or PCA are also available under the *Distance Plot Method* sidebar tab. This combination of input data and statistical method provides us with a two-dimensional representation of the cohort where the distance between the observations approximates the true similarity of the patient descriptors. In terms of the visualization choice, the use of a scatter plot also allowed us to incorporate highlighting techniques into later interactive elements that are intuitive and easy to interpret by broad range of users.

Element (2) This element gives an initial glimpse into the clustering potential of the patient cohort. Its main goal is to identify outliers in the dataset and patient groups that form a core for specific clusters even when the number of clusters is increased. Two main alternatives were considered for this task: a *classical dendrogram* and a *Sankey* plot. A dendrogram, although a more conventional tool for this task, is mainly used in combination with hierarchical clustering, where a height parameter—representing the distance—is used as an indication at which point certain clusters are joined. However, such a height measure is specific to hierarchical clustering and is not compatible with other clustering methods evaluated in this work. An additional drawback of a dendrogram is that it only works when the entire hierarchy from 1 to n clusters is visualized, since the size of each cluster is only visible by tracing back each node to its lower-level components. On the other hand, a Sankey plot gives a clear visual indication of cluster sizes even when we limit the analysis to a specific range of clusters considered, appropriate for our use case. Although the individual patients are not directly visible in this plot, we have enhanced the visualization to display the list of patients upon hovering over specific nodes.

Element (3) This element serves as a validity measure for the comparison of certain numbers of clusters. The goal of this visualization is to extend the visual inspection of specific clustering results with statistical measures that give further indication of clustering quality. For this task, we have implemented an elbow plot [Tho53] in the form of a line graph. This type of plot is commonly used in use cases where clustering techniques are employed, and has also been widely used in previous related work. Therefore, it was considered a suitable choice for a wide range of target users. As alternative solutions or further extensions of this part, one could rely on statistical measures such as the

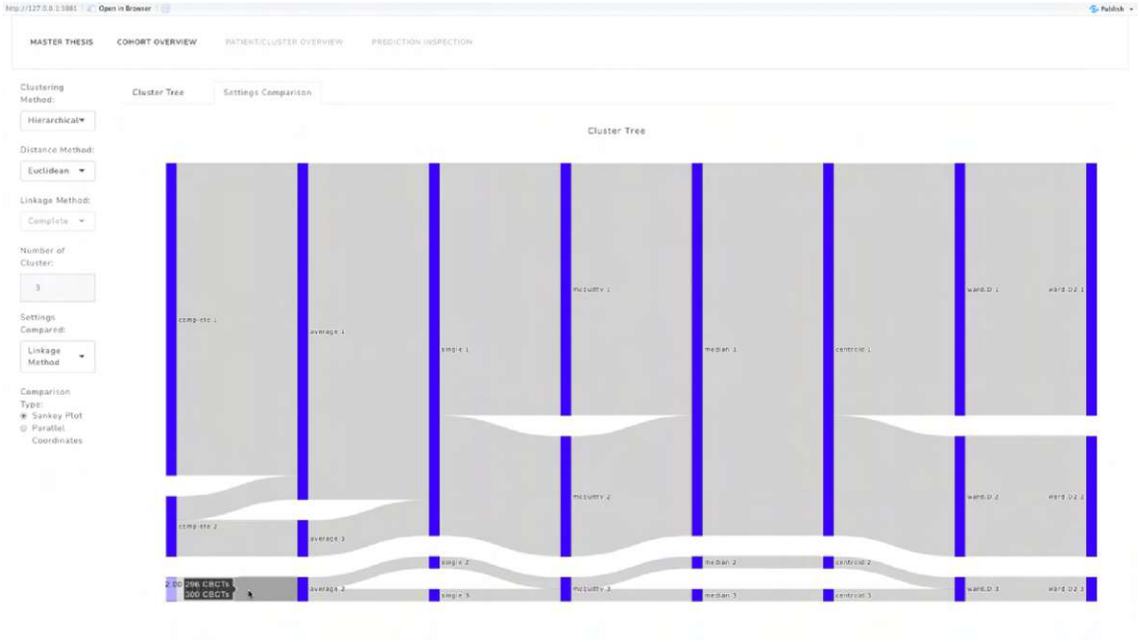
Silhouette diagram [KR90a] or the Gap statistics [TWH01]. However, these are less widely used and require deeper statistical knowledge from the target users.

In Figure 4.20, we relied on hierarchical clustering under Euclidean distance and complete linkage, settings that have also been used by Furmanová et al. [FMCM+21] in their publication. We can already see in the scatterplot (Element 1), that under these clustering settings, two observations tend to be separated from the rest of the cohort as possible outliers. This is also illustrated by the Sankey plot (Element 2), where we can see how the cohort is split into an increasing number of clusters. It can be seen how the two outliers are split into their own cluster in the first split and into two separate clusters when the number of clusters is increased to 6 or more. Finally, the elbow plot (Element 3) serves as a guide to illustrate the improvement in cluster compactness with the number of clusters increasing. It can be seen that the steepest decline in the within cluster sum of squares is observed at the first split, which isolates the outliers, after which the decline becomes more linear. Given our prediction workflow and the general use case, selecting 3 or 4 clusters under these settings would be an optimal choice. After this point, we can also see in the cluster tree that further splits lead primarily to an increased number of clusters with only a few observations, leaving the largest cluster of *Cluster 1* intact.

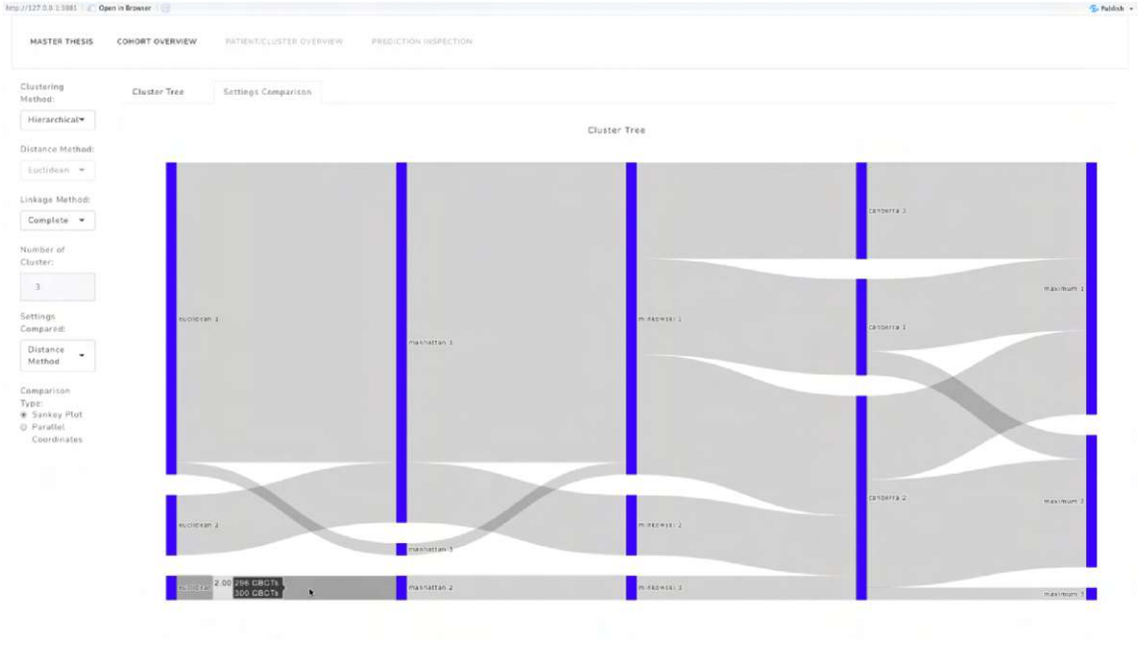
In addition to the number of clusters, we are also interested in the effects of using different linkage and distance methods in hierarchical clustering. For this purpose, we have created the *Setting Comparison* sub-page. In Figure 4.21 (a), we show the page when used for the comparison of different linkage methods, while in Figure 4.21 (b) for the analysis of various distance methods. In Figure 4.21 (a), we can see that linkage methods such as *single*, *median*, and *centroid* linkage tend to separate individual observations as outliers, while others result in more balanced cluster sizes. For the linkage methods, we can observe that the *Euclidean*, *Manhattan*, and *Minkowski* distance (with exponent $p = 3$)—which in fact differ only in the exponent used to calculate the distance between observations—all yield similar results, while methods such as the Canberra or maximum distance carry more randomness because their calculation involves specific aspects of the shape descriptors (see Section 4.1.3).

Overall, this page serves as an initial overview of the patient cohort with a particular combination of clustering settings chosen by the user. Using it, we were able to visually identify that there are two possible outliers in the cohort that are also clearly identified by the clustering algorithm. Using the information presented, we were able to determine an appropriate number of clusters for our use case. If we were also interested in alternative clustering approaches, we could switch to other methods for exploration purposes, or specifically for hierarchical clustering, we could access the *Settings Comparison* sub-page to gain insights into the effects of alternative settings. Such insights are of key interest for us when investigating research questions **RQ 2.1** and **RQ 2.2**. Furthermore, by comparing the results under different subsets of the patient cohort, this page facilities the evaluation of **RQ 2.3** as well.

Figure 4.20: 1st Page—Cohort Overview: Focus on patient cohort composition, identifying an optimal number of clusters and possible outliers



(a) Comparison of different linkage methods for hierarchical clustering



(b) Comparison of different distance methods for hierarchical clustering

Figure 4.21: 1st Page—Cohort Overview: Settings Comparison sub-page

4.3.3 2nd Page—Patient/Cluster Overview

The second page, shown in Figure 4.22, allows for an in-depth analysis of the various clustering results and their respective predictive performance. Since the analysis for this task involves a large number of possible features, the most important consideration during the design of this page was to allow the user a high degree of customization.

Element (1) This parallel coordinates plot is the main feature on this dashboard page. This representation allows the user to customize the input data and the dimensions of the analysis depending on the task at hand. To support this, the features can be divided into two categories by the user: fixed features, and variable ones. For the fixed features a unique value must be selected as a filtering setting for the data included in the analysis. Variable features are used as axes for the parallel plot. Furthermore, the order of the axes in the parallel coordinates plot can be modified by the user, thus enabling a convenient inspection of the correlation between any pairs of variables.

Element (2) This barplot highlights a very specific point of interest, which is the frequency of individual cohort patients clustered together with the patient of interest. This summary takes into account all the different outcomes for the settings selected by the user and gives an indication of the similarity of each patient to the patient of interest. It is clear that the patients of interest are always clustered together with themselves. Of key interest here is whether certain patients are also frequently clustered together with them. This information could help identify patients who are always important for the prediction regardless of the clustering settings chosen. Future evaluations could also consider hand-selected clusters, where such information could aid the selection process.

Element (3) The third element of this page is a data table that contains all the predictions for the settings selected by the user. This gives the user feedback and insight into what exact data the current analysis and visualizations are based on. However, its primary function is to identify a smaller number of observations when the parallel plot is filtered for specific observations—for example, those with the best predictive performance. Furthermore, an additional interactive feature allows the user to click on a specific observation in the data table, thus getting redirected to the third page of the dashboard, which displays detailed information about that prediction.

The second page involves a more patient-centered analysis and requires a *Patient of Interest* as input setting. In the context of our work, we select one of the patients from the input cohort for this purpose, providing only a limited number of *Timesteps* to simulate the scenario of a newly incoming patient with incomplete data. In addition, the user has the option to select fixed and non-fixed variables to be used for the axes in the parallel coordinates plot. In Figure 4.22 we examine the predictive performance at an *ISO* level of 0.5 under different hierarchical clustering parameterizations for the organ of the prostate. We then proceed to focus on the best and worst performing settings, by applying filters in the the parallel coordinates plot (Element (1)). Here we can see that

both predictions were generated under binary linkage using different linkage methods. In addition, in Element (2), we can identify the frequency of patients included for the (filtered) predictions in the parallel coordinates plot. We can see that 13 patients are included for both predictions, while 10 others are included for only one of each. While it is not apparent in Figure 4.22 under what exact settings each patient is included, such information could be made visible by rearranging the axes, which is implemented as a drag-and-drop feature. However, as can be seen in the parallel coordinates plot, although we can successfully identify certain well-performing settings using this dashboard, all 48 setting combinations achieve a performance within a range of 5 percent absolute difference. The insight presented in this page primarily address **RQ 1.3**, while also providing additional quantitative insights about different clustering settings, this way addressing further aspects of **RQ 2.1-2.2**.

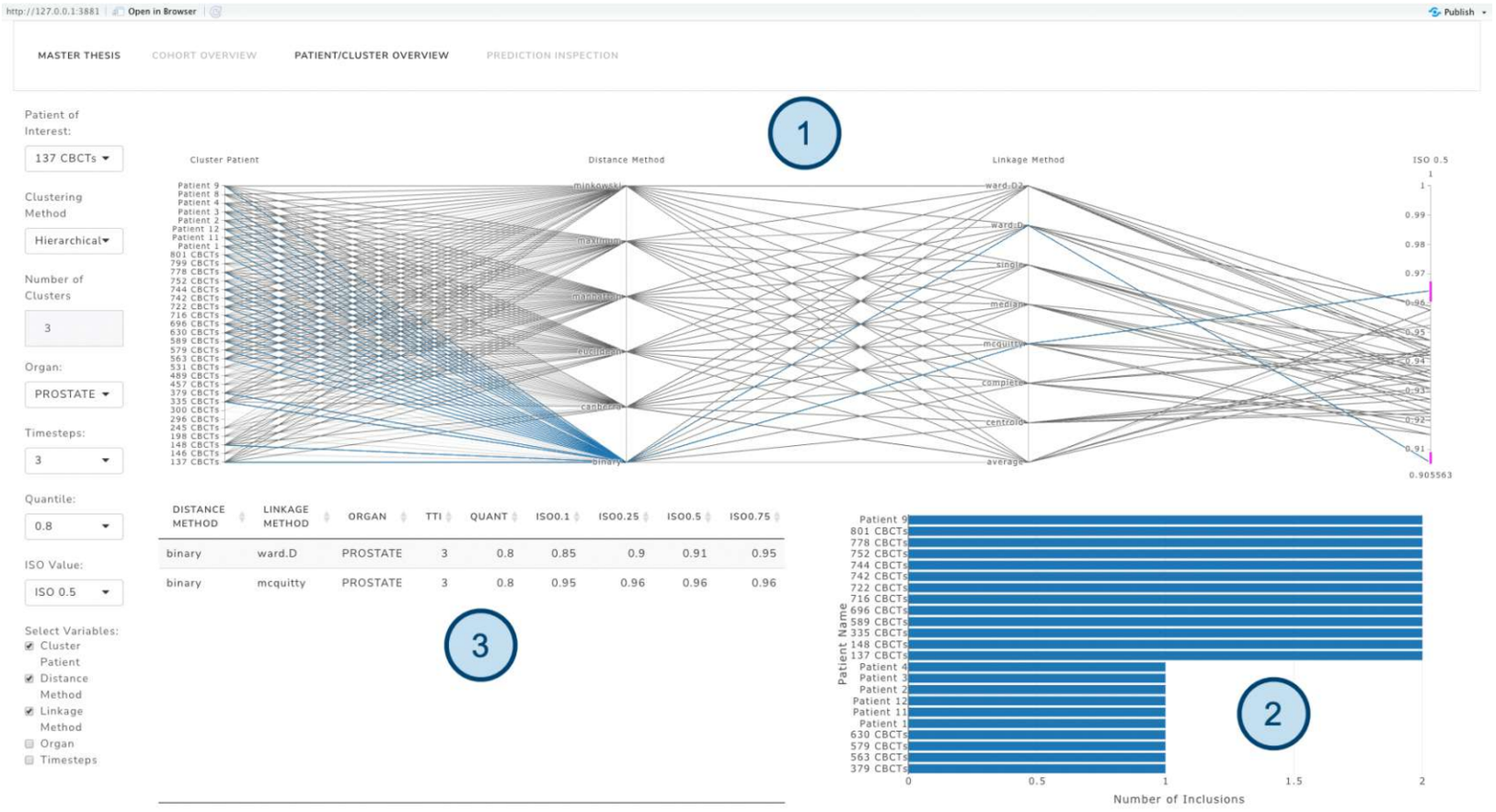


Figure 4.22: 2nd Page—Patient/Cluster Overview: Focus on the exploration of the predictive performance enabled by different settings

4.3.4 3rd Page—Prediction Inspection

The third page, shown in Figure 4.23, is dedicated to exploring the details of the predictions. For this purpose, we have implemented four key visualizations. While the first three elements are specific to the individual predictions, Element (4) extends the analysis and includes information about performance change when modifying the timestep or quantile setting, thus providing further context.

Element (1) This radial plot [HGP99] focuses on the cluster patients involved in the prediction workflow for the new patient. Here, each point is a patient, with the patient of interest in the center. In the surrounding area, the cluster patients are distributed in a circle, with their distance from the patient of interest representing their descriptor similarity. This gives a first indication of the size and composition of the cluster patients used for the predictions. By hovering over a particular patient, we can view the identifier of the underlying patient. At the same time, the same patient is also highlighted in Element (2).

Element (2) The purpose of this element is to summarize the percentage of overlap between the patient of interest and each cluster patient using a bar plot. It goes without saying that an important goal of an optimal clustering solution is to identify the patients with the greatest organ overlap as the most similar. In this regard, it is expected that the closer the patients are located to the new patient in Element (1), the higher they should be ranked in Element (2).

Element (3) This visualization takes the actual shape predictions and compares them to the target shape in three dimensions using a 3D scatterplot-like grid. The main consideration here was to allow the user to examine the actual organ shape predictions and not just the resulting performance metrics. This could help for example to identify specific regions of the shapes, where the deviation between the prediction and the target is visibly worse than elsewhere.

Element (4) The final element on this page provides a more general overview of the prediction performance for the patient of interest. While a single prediction can only be analyzed for a specific combination of timestep and quantile settings, it is important to have an overview of how the performance would change along these two dimensions. It should be noted that the same could be true for different ISO cut-off values, but as we will argue in Section 5.1.1, we can define optimal settings for the cut-off value that rarely require change except for certain exploratory use cases.

In the example shown in Figure 4.23, the radial plot (Element (1)) provides a concise visual overview about the size of the cluster and its compactness. We can see that most patients have approximately a similar distance to the patient of interest, with two patients having a higher distance. In terms of the general trends, we expect this ranking to align with the ranking shown in the barplot in Element (2) as well. This

visualization shows the overlap of the patient of interest with the other patients in the cluster, with fill colors depicting it on the level of individual organs. However, the barplot is based on the Dice coefficient between the organs, while the radial plot is based on the distance between the shape descriptors which also include additional aspects such as the standard deviation at specific positions, resulting in no perfect match in their ranking. In addition, under the *Full Cohort* sub-page it would be also possible to extend the contents of these two visualizations to include the entire cohort to inspect the separation of the cluster patients from the remaining ones in the cohort. Next, in Element (3) we can make a three-dimensional comparison of the target and predicted shapes for an visual inspection of their overlap, helping also to identify critical areas where the the prediction quality is worse, comparably to VAPOR [FGM+20] and PREVIS [FMCM+21]. Finally, in Element (4), a line chart gives an indication about the importance of the two remaining variables unaddressed, namely the *Timesteps* included for the predictions and the *Quantile* settings selected. In Figure 4.23 we can identify, that the two quantiles of 0 and 1, including the most extreme shape variations are the most challenging to accurately model in the predictions. All other quantile settings tend to perform similarly, with the prediction performance slowly increasing as we include more and more timesteps, with the highest observable performance increases visible for the inclusion of the first 5 CT scans. Such insights, in particular about the effects of the number of timesteps included and the quantile settings, help us to answer **RQ 1.3**. Finally, while **RQ 1.1** and **RQ 1.2** can be primarily explored by using modified input data sets—e.g., ones based on different shape description methods or modified CT scans to introduce missing slices—we anticipate the most visible impact of these modifications to be visible on this dashboard page. In particular, the overlap comparisons in Element (2) and Element (3) are of key interest for these research questions.

Figure 4.23: 3rd Page—Prediction Inspection: Focus on individual organ shape predictions, the patients included for their generation

4.3.5 Interactivity

A key feature of the developed visual analytics application is the interactivity of the components. In *Page 1* (see figure 4.20), clusters of patients are represented as nodes in the Sankey diagram. Hovering over individual nodes lists the underlying patients in the cluster to provide an overview. At the same time, the same patients are also highlighted in the patient similarity scatter plot. The latter feature is also implemented when the hovering over specific points in the Elbow plot. These features ensure that the cluster composition remains transparent to the user. In *Page 2* (see figure 4.22), the emphasis is on the parallel coordinates plot. Besides the option to manually select the variables that are used as axes in the visualization, the order of the axes can be adjusted using a drag-and-drop mechanism. Furthermore, the use of filters on the axes enables a more focused analysis. Setting such filters automatically updates all other elements on the page as well. This includes the data table, where individual observations can be selected for detailed inspection by clicking on them. Upon this action, the user is redirected to the final dashboard page. In *Page 3* (see figure 4.23), both the radial and barchart visualize a similarity ranking of the patients. However, these similarity measures are calculated based on different aspects. To facilitate the comparison of the rankings, when hovering over a patient in the radial chart highlights the corresponding bar in the bar chart.

CHAPTER 5

Results

In Section 4.3 we have already seen examples of how the developed visual analytics interface facilitates a qualitative investigation of the prediction workflow with respect to individual patients. In this chapter, we take a more quantitative focus and present the results of our research, conducted according to the statistical evaluation approach outlined in Section 4.2. To address our research questions presented in Section 1.2, here we are interested in general insights, for which we summarized our findings in the form of various statistical simulations. In these we frequently compare, rank, and measure deviations between alternative methods evaluated.

5.1 Shape Descriptors

5.1.1 Research Question 1.1

RQ 1.1: *What are the effects of using different shape descriptors?*

To evaluate the ability of different shape descriptors to reconstruct the original input shape, we first focus on the probabilistic shape description method used in PREVIS (see Section 4.1.1 for details). Figure 5.1 presents a comparison of the up-sampled shape descriptors using various ISO cut-off values and their respective reconstruction performance measured by the Dice coefficient. Each line in this visualization represents a single patient, with the results calculated as an average performance of the available timesteps for a given organ.

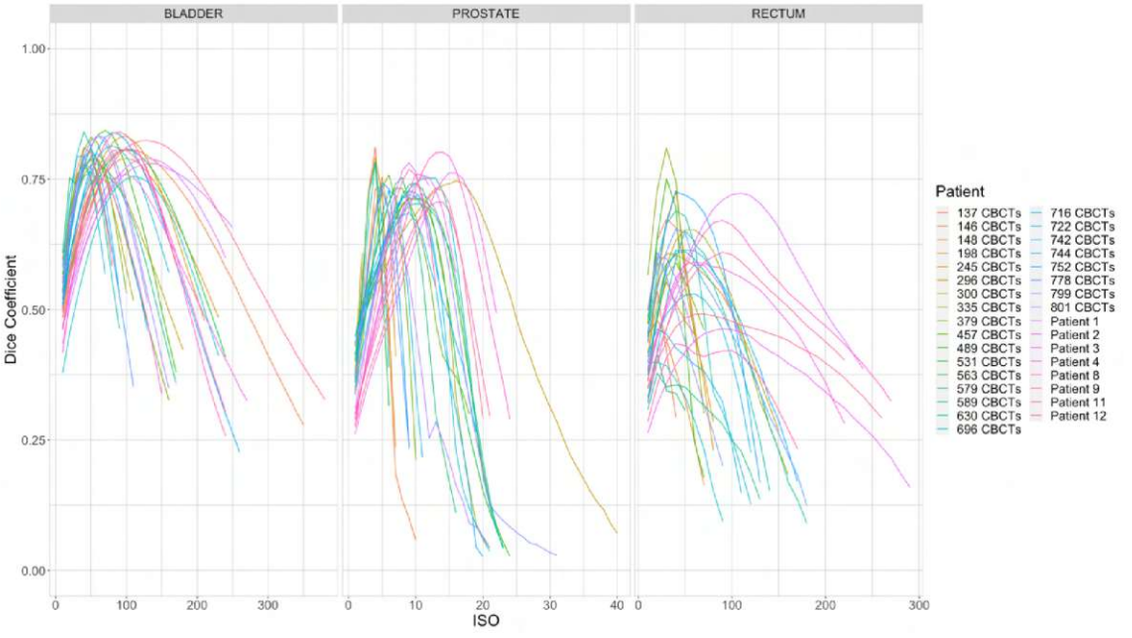


Figure 5.1: PREVIS descriptors: reconstruction performance measured by the Dice coefficient at different ISO cut-off values

Taking a look at Figure 5.1, it becomes visible that different patients require significantly different ISO value settings to obtain the best reconstruction of their shape descriptor. It can also be seen that the reconstruction overlap barely reaches 80%, even at the best possible settings. The rectum in particular tends to provide the worst results, as it can be seen by several patients peaking below an overlap of 50% in Figure 5.1. However, this was expected, due to the more irregular and non-spherical shape of the rectum, compared to the other two organs. It is worth noting, that this visualization presents the average performance across different timesteps of the same patient. This approach is based on the assumption that the descriptors of individual patients will achieve similar reconstruction accuracies across different time steps for identical ISO settings. However, if we inspect a single cut-off value in detail, we can see that there are in fact noticeable deviations across timesteps. This is presented in Figure 5.2. In this visualization the cut-off value for each organ has been chosen, as the one yielding the best results on average, thus simulating the optimal setting for a cohort wide analysis. Then the reconstruction performance across different timesteps was summarized on patient level using boxplots. It can be clearly seen that not only does the reconstruction accuracy differ significantly between different patients, but the boxplots also indicate that the quality of the reconstructions vary greatly even for different time steps of the same patient. In case of the prostate, this analysis also highlights one of the issues with the irregular range of the descriptor values. While the optimal cut-off value has been selected based on all the prostate descriptors in the cohort, some descriptors might not even contain values above the selected cut-off level, leading to empty boxplots (highlighted by a red square in Figure 5.2).

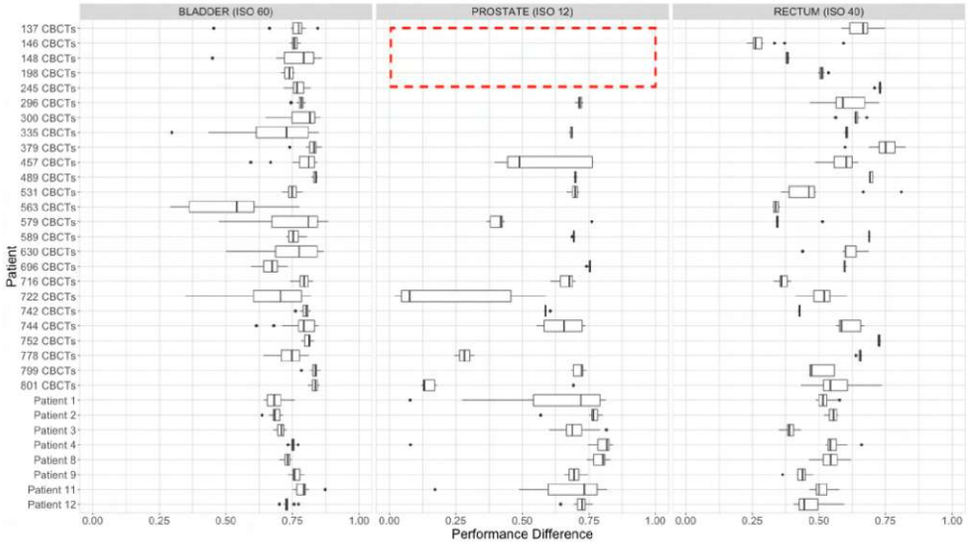


Figure 5.2: PREVIS reconstruction performance deviation across different timesteps of individual patients

Performing the same analysis for the resolution-based descriptors (see Section 4.1.1 for details), we get much better results, as shown in Figure 5.3. Apart from more consistent patterns, this method also achieves an overall better reconstruction performance. The peak performance across all three organ types is achieved by a cut-off value of 0.5, or 50% if speaking in terms of probabilities. This aligns with our expectations, as this threshold describes the limit above which the *majority* of a given target region is part of the underlying organ.

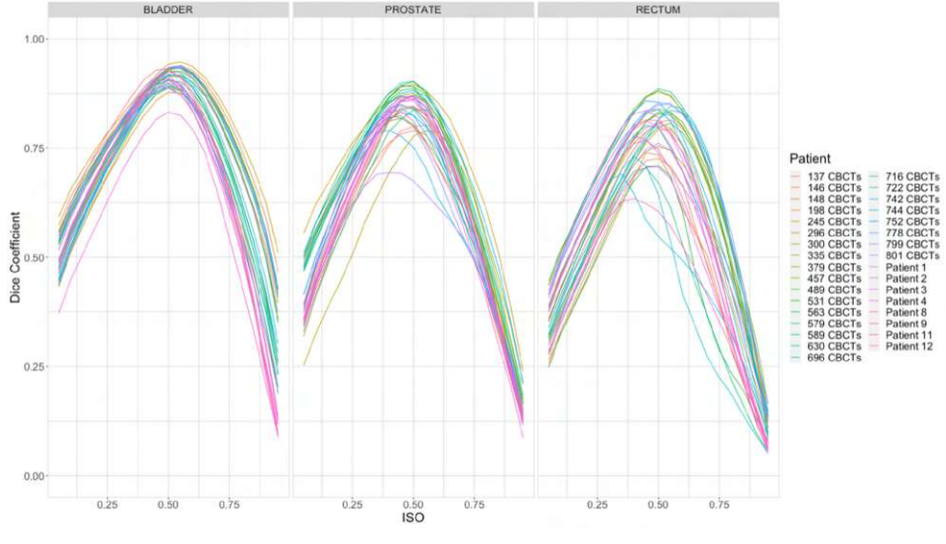


Figure 5.3: Resolution-based descriptors (15mm grid resolution): reconstruction performance measured by the Dice coefficient at different ISO cut-off values

Using the same approach as before to examine a single cut-off level across multiple timesteps, Figure 5.4 shows also much more consistent results. Not only is the reconstruction performance more similar across different patients, but there is much less deviation within the same patients across different timesteps. This can be seen when comparing the interquartile range covered by the boxplots in Figure 5.2 and Figure 5.4. The most notable deviations can be observed for the rectum.

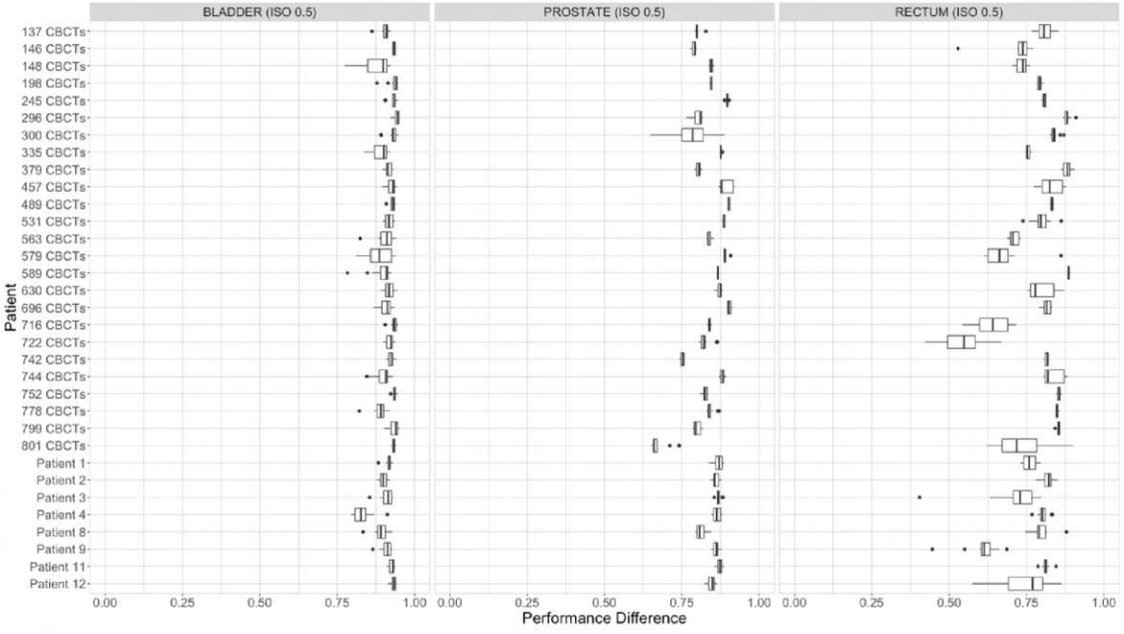


Figure 5.4: Resolution-based descriptor reconstruction performance deviation across different timesteps of individual patients

The rectum tends to yield the worst reconstruction performance, due to its irregular shape, which is not suitable for a description in terms of the large voxels, defined by the descriptor grid. Therefore, opting for smaller grid dimensions, thus increasing the overall number of target points, might improve the quality of the descriptors and their reconstructions. Figure 5.5 demonstrates this by decreasing the dimensions of the grid from 15 to 10 millimeters. Compared to Figure 5.3, it can be seen that the reconstruction performance has increased for all three organs. In addition, the rectum now shows more consistent patterns, without large differences between patients. However, it is worth noting that this improved performance is a result of the larger shape descriptors, which in turn are associated with an increased computational cost.

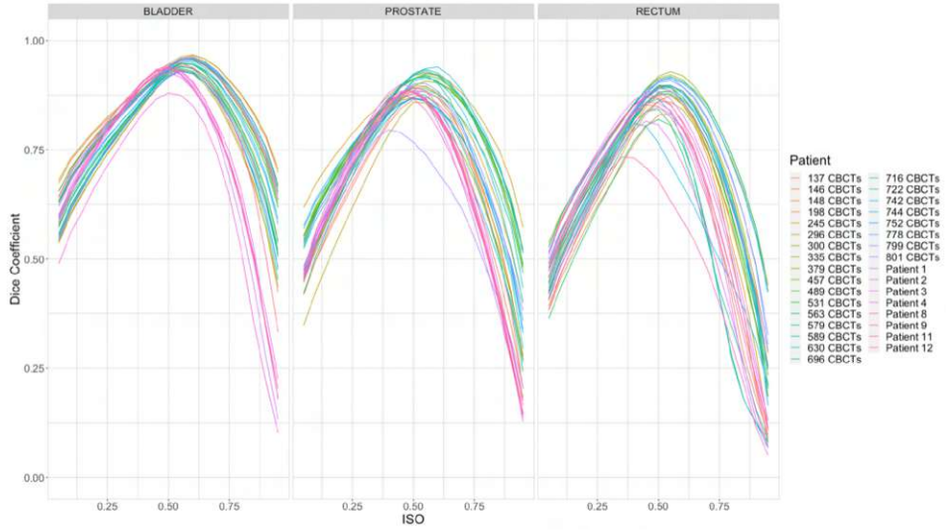


Figure 5.5: Resolution-based descriptors (10mm grid resolution): reconstruction performance measured by the Dice coefficient at different ISO cut-off values

The analysis presented in this section has highlighted several problems with the shape description approach used in the original implementation of PREVIS and the results presented by Furmanová et al. [FMCM+21]. The most critical issue is that the use of a uniform cut-off value is not appropriate for the shape descriptors used in their analysis. In fact, using cut-off values in the range of 0 to 1 while neglecting the range of the actual descriptor values leads to inaccurate representations for the initial input shapes. Moreover, while these representations were still compatible inputs for the prediction workflow, the quality of the resulting predictions was only evaluated at the shape descriptor level, leading to biased results in terms of the actual organ variability prediction performance. The resolution-based descriptors proposed in Section 4.1.1, provide an alternative solution with more accurate and reliable representations for the organs and will be used as a basis for later evaluations presented in this work. Overall, we conclude for **RQ 1.1**, that the choices made for the shape descriptors can have a significant impact on their capability to accurately represent the underlying organs, and thus affect the entire prediction workflow. Therefore, an appropriate selection of the shape descriptor methods and settings employed is a critical part of the approach and serves as a basis for any further analysis.

5.1.2 Research Question 1.2

RQ 1.2: *What are the effects of introducing noise to the input data? How sensitive are various settings to inaccurate input data?*

An important insight from **RQ 1.1** is the relationship between the input data and the shape descriptors. As described in Section 4.2.2 the quality of the shape descriptors does not only depend on the appropriate choice of shape descriptor method, but clearly also on the quality of the underlying input data. Therefore in this research question we focus

on a frequent issue affecting the quality of the input data, namely missing slices in the organ annotations. In our patient cohort, a missing slice in the organ annotation can be for example also observed for the rectum of patient *146 CBCTs* in the planning timestep.

To study the effects of missing slices on the derived shape descriptors, we conduct a two-stage evaluation. In the first stage, we examine the effects of single missing CT slices at random positions. To do this, we randomly exclude individual CT slices, recalculate the shape descriptors, and then compare the reconstruction capability of these descriptors with the original ones. In the second stage, we examine the effects of multiple missing CT slices on the quality of shape descriptors. As the percentage of missing slices increases, we expect to see a corresponding decline in the quality of the shape descriptors. To assess the impact of missing slices this, we will iteratively remove neighboring slices, gradually increasing the amount of missing information.

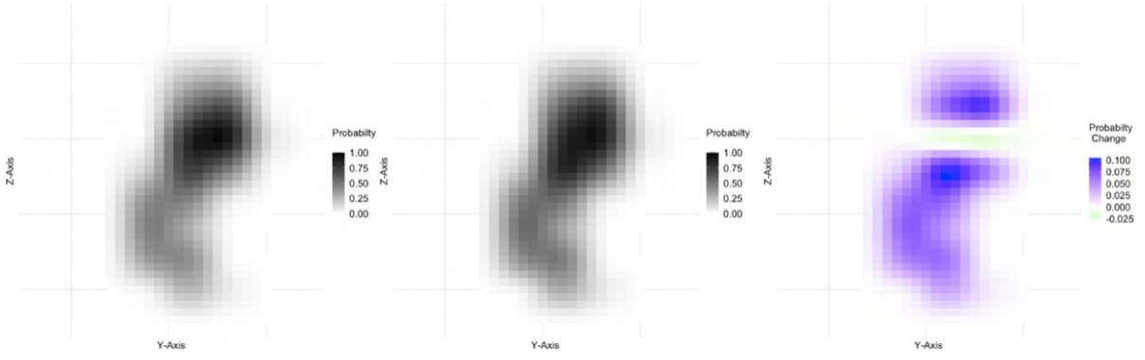
Individual Missing Slices

In line with the previously proposed workflow, we have started our analysis by excluding a randomly selected slice in each organ annotation along the dimension of acquisition—denoted as the z-axis in this section—and recalculated a new shape descriptor for each organ. Due to the nature of the input data, the missing slices were selected for each organ individually, resulting in non-corresponding slices across different organs in the CT scans. However, as our analysis is also conducted on the level of individual organs, this does not impact the results presented in this section. This step generated a modified shape descriptor dataset with 433 observations for each organ. Next, we have upsampled the shape descriptors and calculated their reconstruction accuracy measured by the Dice coefficient. Finally, we have compared these results to the Dice coefficients yielded by the shape descriptors based on the full CT scan. A summary for the observed deviations can be seen in Table 5.1.

Organ	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
BLADDER	-0.03585	-0.00573	-0.00078	-0.00220	0.00252	0.01934
PROSTATE	-0.06407	-0.01101	-0.00489	-0.00577	0.00072	0.05716
RECTUM	-0.14588	-0.01486	-0.00446	-0.00564	0.00319	0.11258

Table 5.1: Summary statistics: reconstruction Dice coefficient decrease with single missing CT slices for the different organs

Upon examination of the mean, median, and 1st and 3rd quartiles, we can see that the difference in reconstruction accuracy is negligible, with a maximum absolute difference of 1.5%. However, the results at the extremes of our observations are more noteworthy. Most interestingly, a relatively large increase in performance was also observed in some cases. To further investigate this unexpected result, we will now present an analysis for the largest positive increase in the Dice coefficient, observed for the rectum of patient *801 CBCTs*, timestep 12, with an absolute increase of 11%. Figure 5.6 highlights the differences in the upsampled shape for this organ using the different shape descriptors.



(a) Using full CT scan (b) With a single missing slice (c) Probability Difference

Figure 5.6: Missing CT slice: upsampled rectum shape comparison (patient *801 CBCTs*, timestep 12)

Figure 5.6 (a) is based on the original shape descriptor, based on the full CT scan. In Figure 5.6 (b), we used the modified data with a missing slice, and in Figure 5.6 (c) we used color coding to highlight the differences between them. While the difference may not be immediately apparent in Figure 5.6 (a) and Figure 5.6 (b), Figure 5.6 (c) shows that the exclusion of a single slice resulted in an absolute change in the probabilities ranging from approximately -2.5% to 10%. The decrease in probabilities is due to the fact that we have excluded a slice exactly from that specific region of the organ. The increase in other parts of the organ can be attributed to the implementation of the upsampling procedure. As described in Section 4.1.1, the final stages of upsampling involve a smoothing step to make the transition between voxels more fine-grained. This can cause some probabilities to decrease as they are affected by the surrounding smaller values, which we address by scaling all probabilities back to a range from 0 to 1. However, in this case, the missing slice was introduced to the very core of the organ, which meant that the smoothing phase resulted in overall smaller values. As a result, during the final scaling, many of the probabilities increased relative to the original settings without missing slices. This phenomenon is generally expected to be infrequent and have a small effect, and requires multiple conditions to be met simultaneously. First, the organ must have a relatively small core region that is easily affected during the smoothing procedure. Second, the missing slice must exactly affect this region of the organ. Finally, the amount of probability increase must push the voxels above the cut-off threshold during the organ reconstruction. As Figure 5.7 shows, this is exactly what happened in this case and the increase in the probabilities pushed a large region of the organ above the cut-off value of 0.5.

(a) Using full CT scan (b) With a single missing slice

Figure 5.7: Missing CT slice: rectum shape reconstruction comparison (patient *801 CBCTs*, timestep 12)

Additionally, we can conclude that this phenomenon is in many ways analogous to using a smaller cut-off value for the organ reconstruction. As discussed in Section 5.1.1, this would be beneficial in some cases, particularly for the rectum, but overall it would decrease the quality of reconstructions if regions with an occupancy probability below 50% were included. Nonetheless, future research could explore other approaches that maintain both the quality of shape descriptors and the interpretability of their values as probabilities.

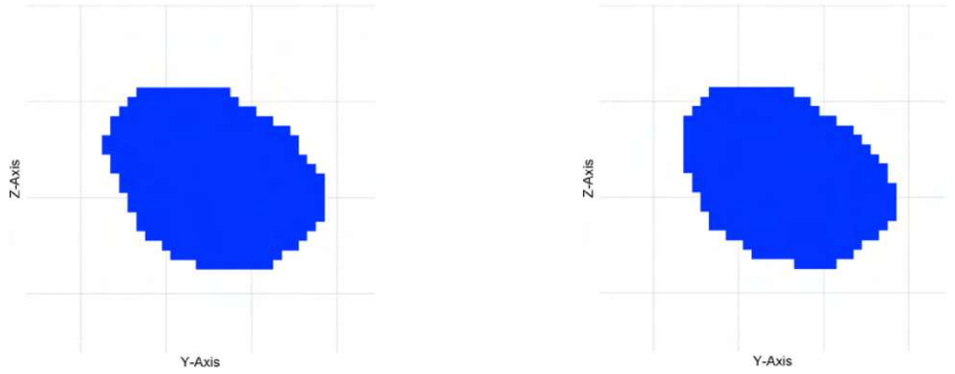
Apart from the outliers described above, for the majority of cases we found that the exclusion of a single slice has only a marginal effect, and is limited to the edges of the organ. To illustrate this, Figure 5.8 (a) shows a sample bladder. Next, we iteratively excluded its source slices, always one at a time, derived a shape descriptor, and then reconstructed the input shape. As expected, in some cases, the missing information resulted in a decrease in the reconstructed occupancy probability, compared to the values provided by the original, unmodified shape descriptor. Figure 5.8 (b) summarizes the largest decrease observed at every position, when taking any possible missing slices into consideration. In other words, this aggregated view shows the worst possible decrease that could be caused by the exclusion of *any* single input slice.

(a) Original occupancy probabilities for organ slice

(b) Highest probability decrease at each position for any single missing slice

Figure 5.8: Aggregated view of impact of individual missing slices on occupancy probabilities (patient *146 CBCTs*, bladder)

When comparing the plots in Figure 5.8, we can see that the deviations are limited to the regions around the edges of the organ and primarily affect voxels with low occupancy probabilities. This also means that the actual organ reconstructions, when using our preferred cut-off value of 0.5, do also not differ in major ways. A comparison can be seen in Figure 5.9. Note again, that this is still an aggregation of all possible single missing slices, to highlight their possible effects on the shape descriptor.



(a) Reconstruction using the original, full data

(b) Reconstruction using modified data with highest observed probability decrease at every position considered

Figure 5.9: Aggregated view of impact of individual missing slices on organ reconstruction using a ISO cut-off value of 0.5 (patient *146 CBCTs*, bladder)

Multiple Missing Slices

In the previous section we have addressed the case of single missing slices in the CT scans. Next we are interested in extending the analysis to multiple missing slices. Furthermore, we want to gradually increase the amount of missing information, preferably excluding random effects to make the settings comparable between patients and timesteps. To this end, we have evaluated the effects of excluding increasingly more neighboring CT scan slices, starting from the center of the organ. We have repeated this for each organ and timestep separately and summarized the gradual performance decrease in Figure 5.10. In this visualization, the reconstruction accuracy achieved by the shape descriptors using the full data was used as a baseline performance.

Figure 5.10: Aggregated reconstruction Dice coefficient decrease with increasing number of neighboring missing slices, summarized on individual organ level

In Figure 5.10 the rectum exhibits the most consistent patterns. This is primarily due to its orientation along the axis of acquisition and its tendency to follow a tube-like shape. As a result, excluding a growing number of neighboring slices effectively removes specific regions of the rectum. In contrast, the effects on the bladder and prostate are more delayed, as the central regions of these organs are more robust and are able to compensate for the missing slices. However, once the number of missing slices reaches a critical point, the quality of the reconstructions decreases rapidly. For the prostate, it can be once again observed that under certain conditions, the exclusion of additional slices can increase probabilities in other regions, resulting in temporary increases in the performance presented in Figure 5.10.

This type of analysis could also be used to identify a breakdown point at which the quality degradation resulting from missing slices is no longer confined to the edges of the organ, but affects its core as well, essentially dividing it in two. While it is beyond the scope of this study to quantitatively determine this measure for all patients and organs, Figure 5.11 provides an example using a 2D view of the center slice of the organ to illustrate the process visually.

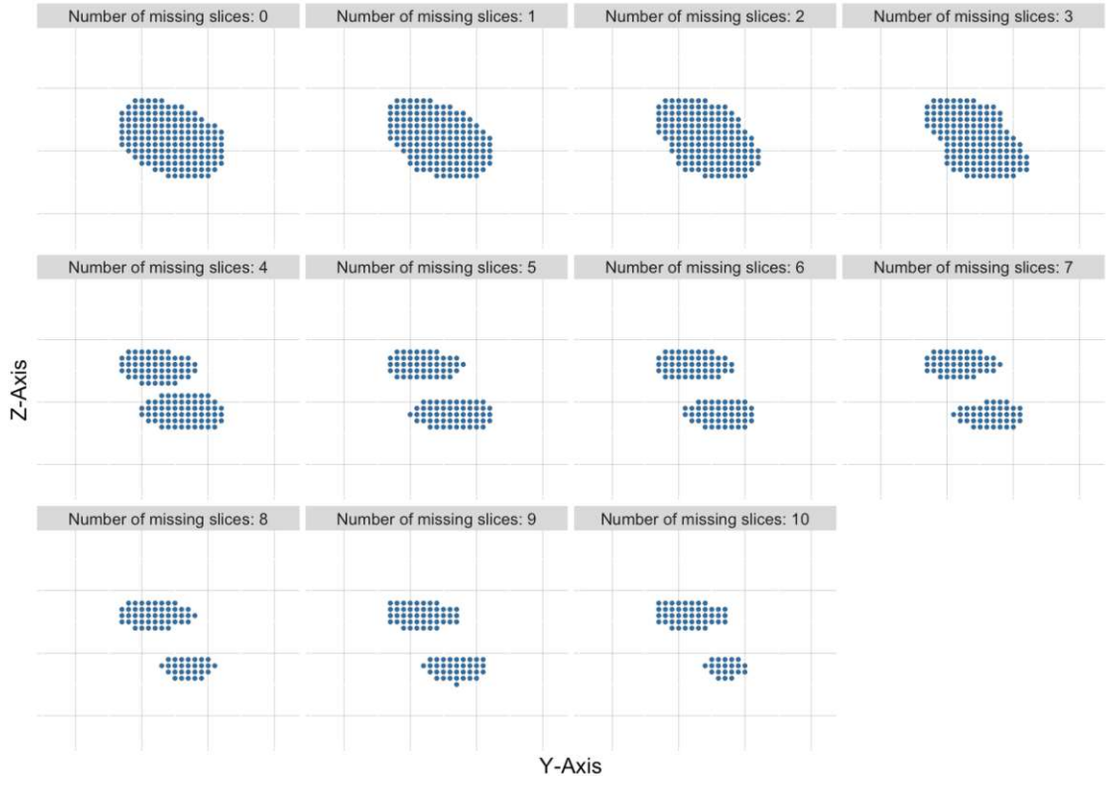


Figure 5.11: Effects of increasing number of neighboring missing slices and breakdown point (patient *146 CBCTs*, bladder)

Based on the insights presented in section, we conclude for **RQ 1.2**, that the proposed shape description method is robust against individual or a low number of missing slices in the organ annotations used as input for the analysis. While we can observe visible impact on the organ reconstructions in certain cases, these effects can be primarily attributed to other parts of the workflow, in particular the upsampling process.

5.1.3 Research Question 1.3

RQ 1.3: *What settings yield the best predictions for new patients with incomplete data? Does it change with increasing information available (i.e., further CT scans)?*

The predictive workflow presented in this work combines two elements to make predictions for new patients with incomplete data. On the one hand, it considers the mean shape of the organs using the available timesteps of the new patient. On the other hand, it uses different quantiles from the shape variation samples generated by a generative model. The distinction between the two parts of a prediction is also important for the results presented in this section.

To illustrate the role of these two components with respect to a single patient of interest, Figure 5.12 summarizes the prediction performance for patient *589 CBCTs* when using

hierarchical clustering with Euclidean distance and complete linkage, which is identical to the settings used in PREVIS. This visualization compares the prediction performance depending on the number of timesteps used for the predictions, with individual boxes distinguishing between different quantiles of added variation.

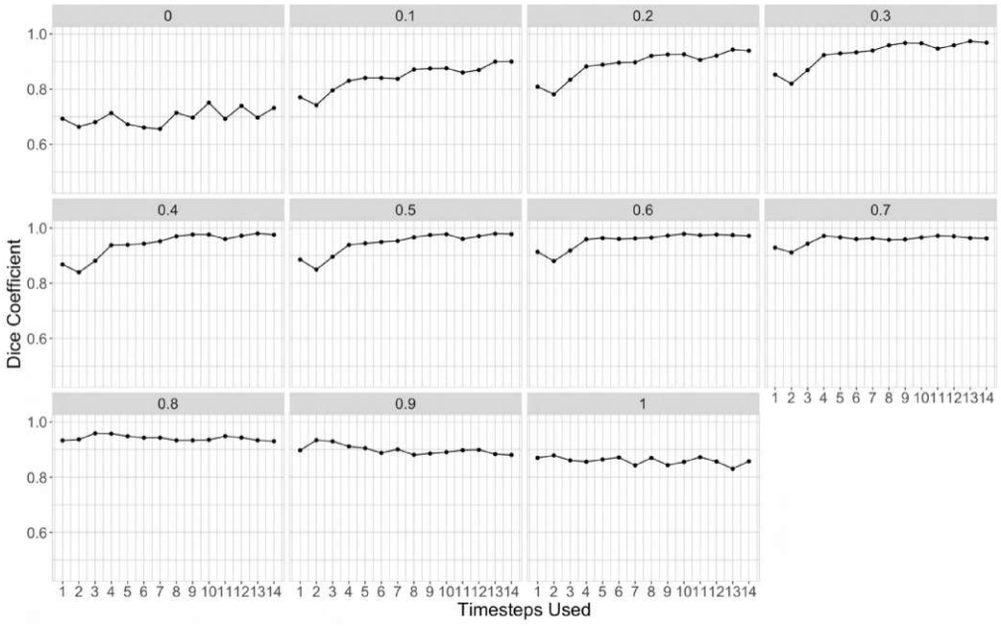


Figure 5.12: Prediction performance for the bladder variation of patient *589 CBCTs* (using hierarchical clustering with euclidean distance and complete linkage with 3 clusters)

As highlighted in Section 4.1.4, the generative model uses normally distributed data to generate sample shape variations. This means that at a quantile of 0.5, the median value of the variations is approximately 0 at all positions. Therefore, under this setting, the prediction is essentially the mean value of the available shape descriptors used to make the predictions. For this setting, Figure 5.12 shows a significant drop in performance when using 2 timesteps and a less significant drop when using 11 timesteps to obtain the prediction. Considering that the prediction consists of the mean of the shape descriptors, we expect that the size or position of the organ at time step 2 and 11 must differ significantly from the other time steps. Since these performance patterns are observed primarily for the lower quantiles, we conclude that the direction of deviation from the mean is especially unfavorable for these settings. This could arise as a result of the the organs having a significantly larger volume at these two time steps. This hypothesis is further explored in Figure 5.13, which shows the patient's bladder volume at each timestep. The red line illustrates how the mean volume changes with each additional new scan. It is also worth noting that large variations such as these have a greater impact on the mean shape if they occur when only a smaller number of scans are available. This can also be seen in Figure 5.13, where the deviation at timestep 2 is much larger than at timestep 11, even though the timesteps are similar in terms of volume.

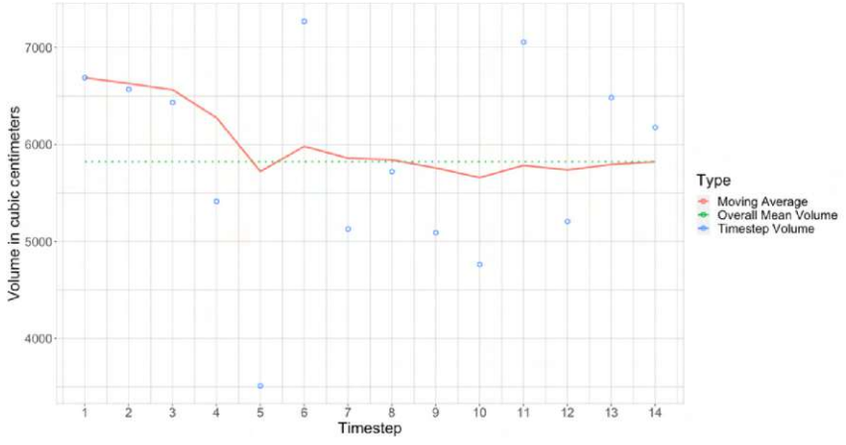


Figure 5.13: Bladder volume analysis of patient *589 CBCTs*

While these results addressed a single clustering setting, Figure 5.14 extends the analysis to multiple possible clustering settings and visualizes the deviations in the performance by using boxplots. Again, using the quantile of 0.5, we can see that the results remain consistent, since the changes in the cluster affect only the generated sample variations and not the mean shape of the new patient. Focusing on other quantile settings, we can generally observe that the more variation we consider, the less accurate the predictions become. In particular, using the quantile of 0, thus shrinking the shape by the most extreme variations leads to the worst results for this patient.

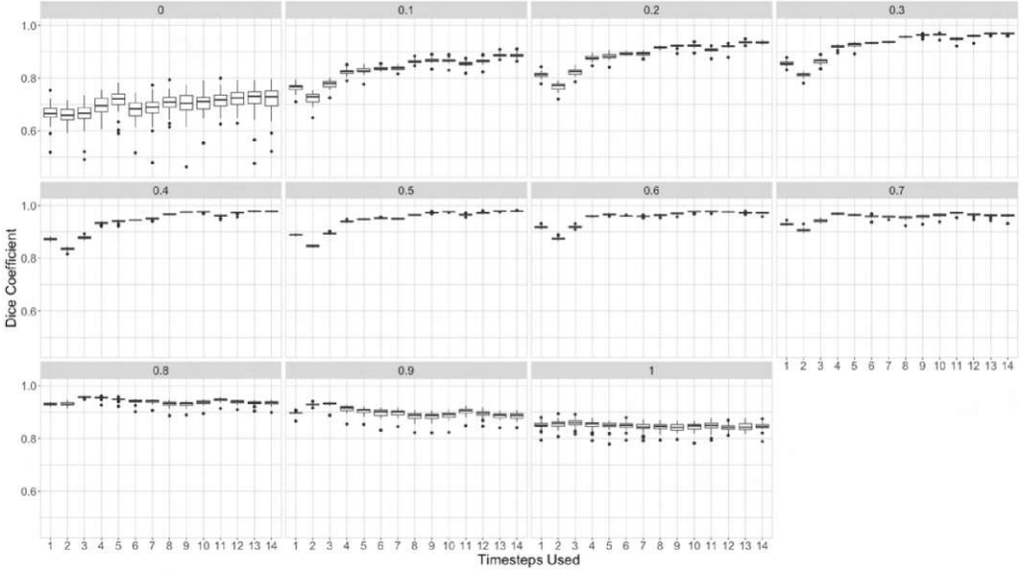


Figure 5.14: Prediction performance for the bladder variation of patient *589 CBCTs* using various clustering settings

We conclude for **RQ 1.3** that the number of time steps included plays an important role in the quality of the predictions, with a higher number of CT scans generally allowing for higher performance. However, outliers in the CT scan series can have a visible impact on the mean shape and quality of the predictions. In terms of quantile settings, we can observe the lowest and most unstable prediction performance for the two quantile settings of 0 and 1, where the most extreme cases of organ shape variation are considered.

5.2 Clustering

5.2.1 Research Question 2.1

RQ 2.1: *What are the effects of using different parameterizations in the clustering (e.g., different similarity measures, different linkage methods)?*

For the topic of clustering, we first focus on hierarchical clustering, which was also the method of choice in the original implementation of PREVIS [FMCM+21]. This approach consists of two important steps, each of which is reliant on a key parameter. First, a distance method is required to compute a distance matrix between each pair of observations in the cohort. Second, a linkage method determines the way different clusters of observations are compared and subsequently merged into larger clusters. The analysis presented in this section addresses these two specific settings and their effects on the resulting clusters.

Distance Method

Focusing on the calculation of a distance matrix, the goal of this step is to compare different observations for their similarity using a specific distance method. The patient descriptors used as input data for this step capture the mean shape and standard deviation at specific positions of the shape descriptors for each individual patient. Of these two aspects, the overlap between the mean shape of two different patients is directly measurable by the Dice coefficient, which allows us to evaluate how well different distance measures correspond to the physical overlap of the organs. A high overlap of the mean shapes is also a prerequisite for making use of the standard deviation part of the descriptor, since it implicitly means that any variation around the organs must occur at the same positions for the two patients. The following analysis summarizes the ability of various distance measures to capture the similarity between the mean shape of different patients. Optimally, a decrease in the overlap between two patient's organs should be accompanied by an increase in their distance. Thus, if we rank the patients in the cohort based on their similarity to a single patient of interest, the ranking should present a steady decrease in the Dice coefficient. By repeating this process for each patient, Figure 5.15 illustrates the mean overlap for the bladder using different distance measures. It also compares the two different types of shape descriptors, centered and non-centered ones.

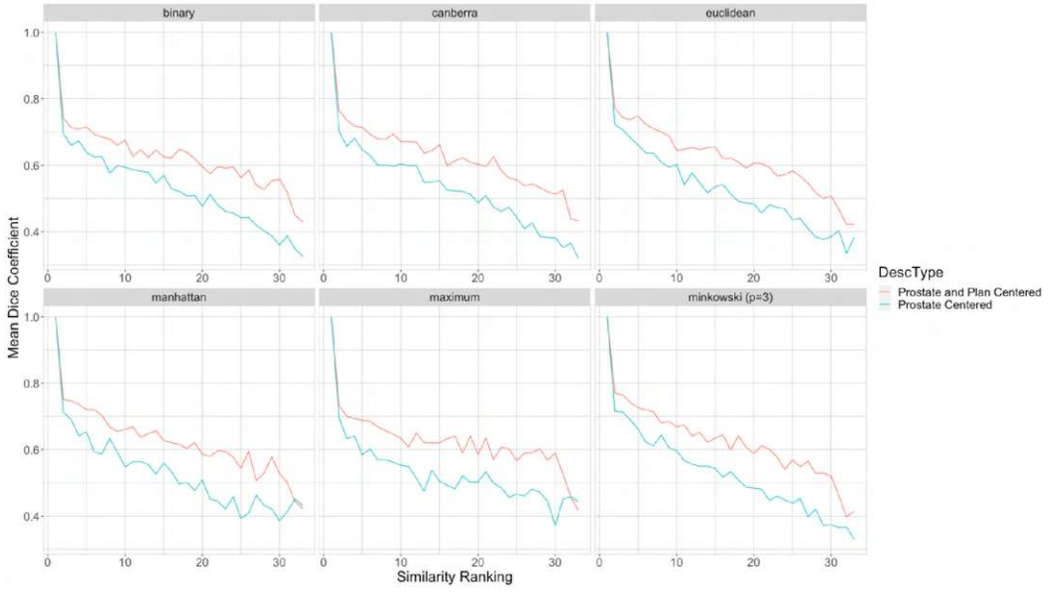


Figure 5.15: Mean overlap of bladders depending on distance ranking

Comparing the different distance methods in Figure 5.15, it is noticeable that the maximum distance measure shows the least optimal patterns. In this setting, a sharp initial decrease in the mean overlap is followed by a fairly constant performance, without the desired steady Dice coefficient decrease as the distance increases. Considering that this approach computes a distance by focusing exclusively on a single position in the shape descriptor with the largest difference, it makes sense that the overlap over the entire organ is neglected. If we focus on other distance measures, we can see more or less similar patterns across them. A further desirable property of an optimal distance measure would be that it does not lead to sudden variations in performance, particularly an increase in the mean overlap. This would imply that the ranking is not optimal and that a larger distance may be associated with an increasing overlap between organs, rather than a decreasing one. This requirement is best met by the Euclidean distance, especially for the prostate and plan centered dataset.

The same analysis can be performed for the other organs of interest as well. Figure 5.16 shows the results for the prostate. In this particular case, we can see minimal differences between the two descriptor types, as all CT scans are prostate centered to avoid any systematic difference between them. In fact, since the additional centering is redundant in this case, the visible differences between the two descriptor types are only present due to the influence of the other organs included in the patient descriptors used for the clustering. Apart from these aspects, we can see similar patterns compared to the bladder. The maximum distance measure provides the least optimal results, while the majority of the other methods meet most of our desired characteristics. Finally, the analysis for the rectum is shown in Figure 5.17. For this organ, all distance methods provide similar results, with all of them showing a relatively small overlap between different patients.

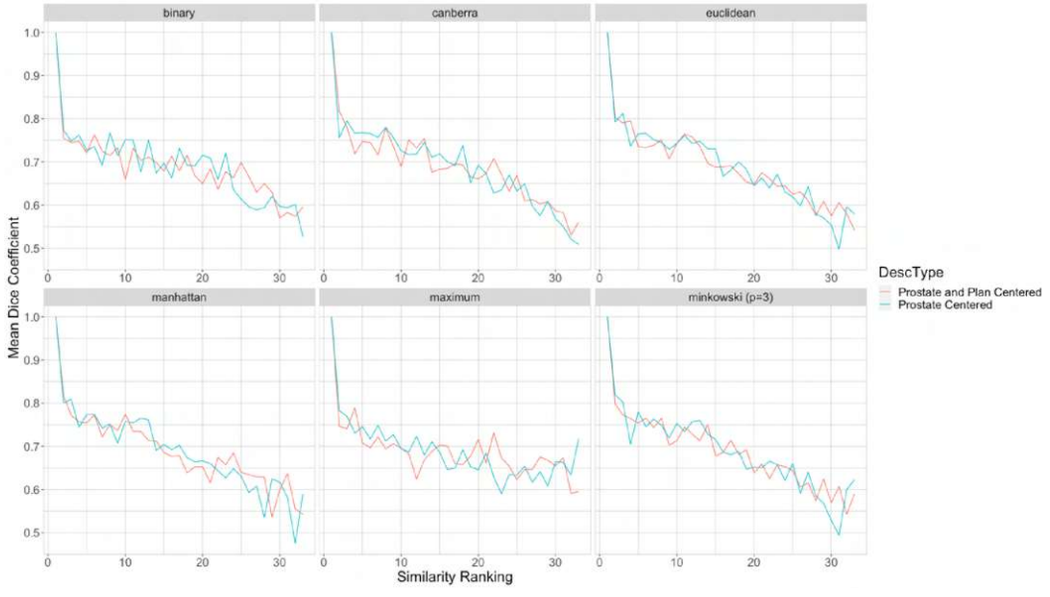


Figure 5.16: Mean overlap of prostates depending on distance ranking

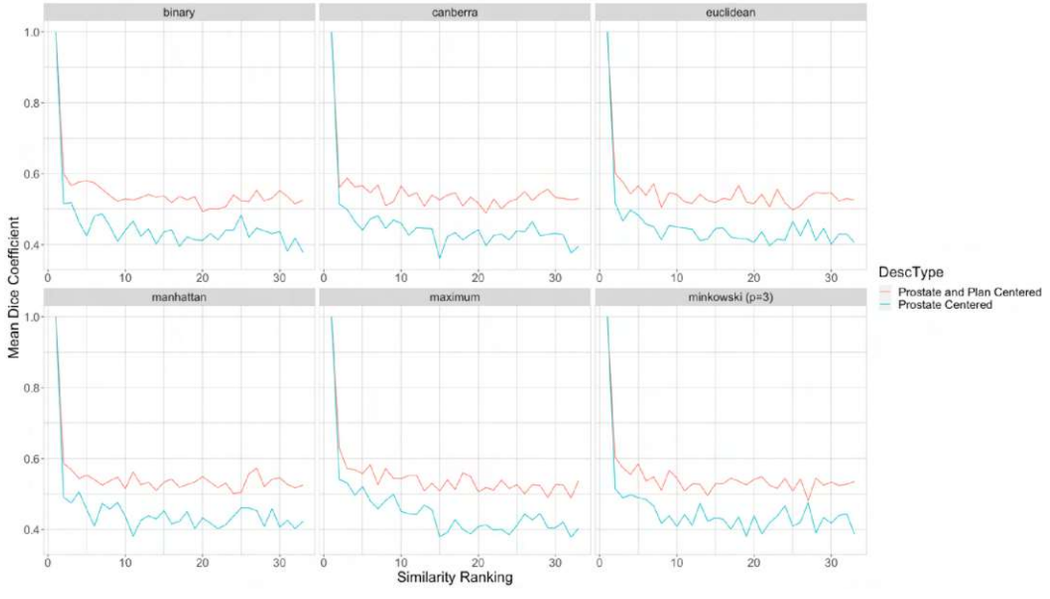


Figure 5.17: Mean overlap of rectums depending on distance ranking

Linkage Method

The linkage method determines how the observations captured in the distance matrix are organized into a hierarchy of clusters. To explore this topic in isolation, we restrict the input distance matrix to the prostate and plan centered descriptors and the Euclidean distance as the distance measure, as this combination has proven to be an optimal setting.

One key aspect of the clustering problem is the selection of the optimal number of clusters. To address this, we visualize and compare the elbow plots for different linkage methods, which is shown in Figure 5.18. In general, the results show that the explained variation increases significantly between the first two clusters. Depending on the linkage method used, the use of 3 or even 4 clusters may be a good choice. The contribution of using more than 4 clusters is only marginal. The only exception to these patterns is the average linkage, which tends to produce more irregular results. While the elbow plot provides a rule of thumb for choosing an appropriate number of clusters, there are other criteria to consider for our use case. The main feature of our prediction workflow is that it generates each prediction based solely on a cluster of the most similar patients. Therefore, the size of this cluster should be appropriate, such that the analysis is limited to a group of patients that nevertheless contains enough patients to collect information for a prediction. In particular, it is not desirable to isolate individual patients as their own clusters, which is another consideration in selecting the optimal linkage method and the number of clusters used. Alternatively, future research could consider an exclusion of certain isolated observations as outliers. However, with the current state of the patient cohort, we are limited to a restricted number of observations where no population-wide distinction between infrequent and true outlier shapes can be made.

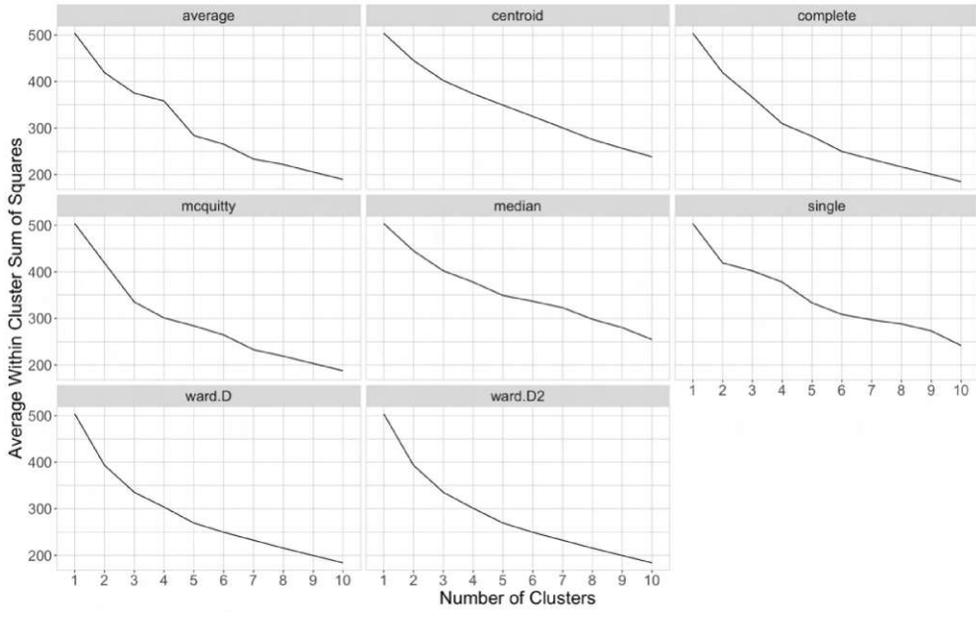


Figure 5.18: Clustering Elbow plots depending on linkage method used

To examine the underlying cluster assignments under different linkage methods, Figure 5.19 and Figure 5.20 visualize the distance matrix and compare the cluster assignment of individual patients using 3 or 4 clusters. For the centroid, median, and single linkage methods, it is noticeable that only one large cluster is present, while all other clusters isolate individual patients in the cohort. Although this can be useful for identifying

potential outliers in the cohort, it is not an optimal setting for our prediction workflow. On the one hand, for the clusters containing only a single patient, the prediction would be based solely on that patient. On the other hand, for the patients in the large cluster, the prediction would be based on almost the entire cohort, except for the outliers. This is expected to be an improvement in itself, but is not the desired goal of this work. For all other linkage methods, the clusters are divided into more equal sized groups of patients, especially for the *complete*, *mcquitty*, and *ward.D/ward.D2* methods. It is worth noting that these qualitative expectations for the clusters are defined according to the core idea behind the prediction workflow. Whether certain clusters with more desirable structure actually lead to better predictions needs to be investigated separately.

Overall, our investigation of **RQ 2.1** has found, that certain distance methods are more capable of measuring and representing the overlap between organs, with a suitable choice being the Euclidean distance. In combination with linkage methods such as the complete linkage, hierarchical clustering provides a fitting clustering for our use case. With the selection of of these settings being identical to the ones used in PREVIS, we have also affirmed the choices made by Furmanová et al. [FMCM+21].

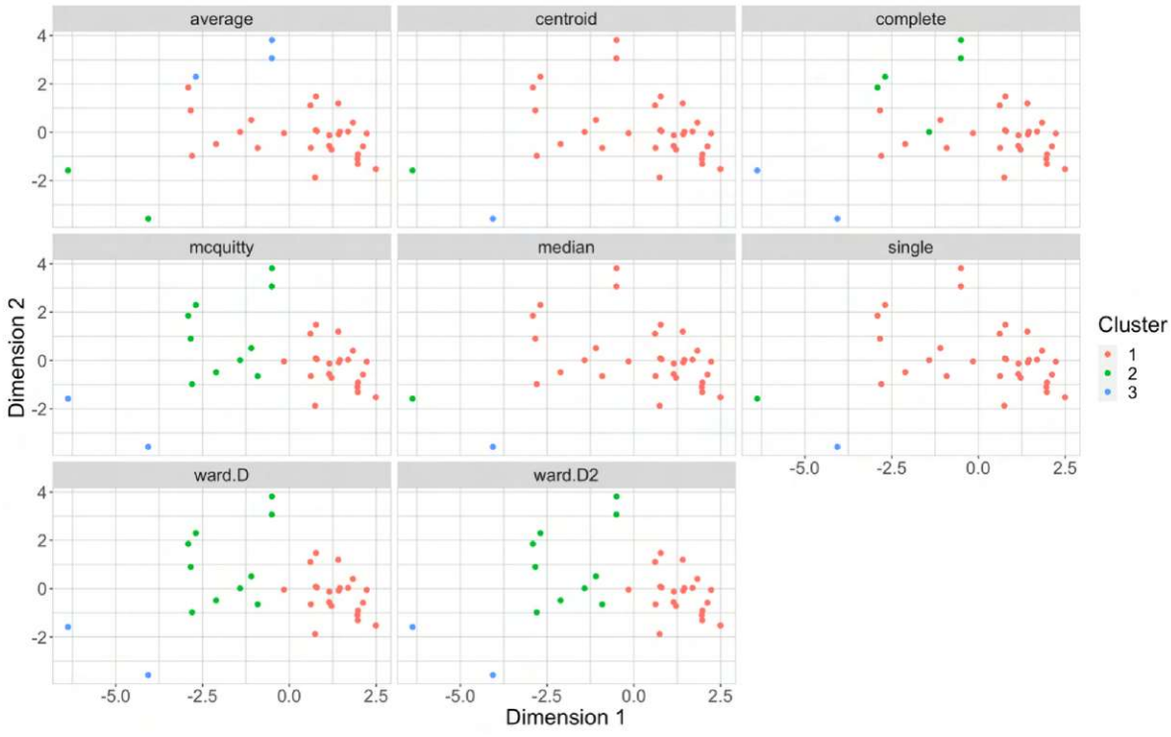


Figure 5.19: Patient assignment for various linkage methods using 3 clusters

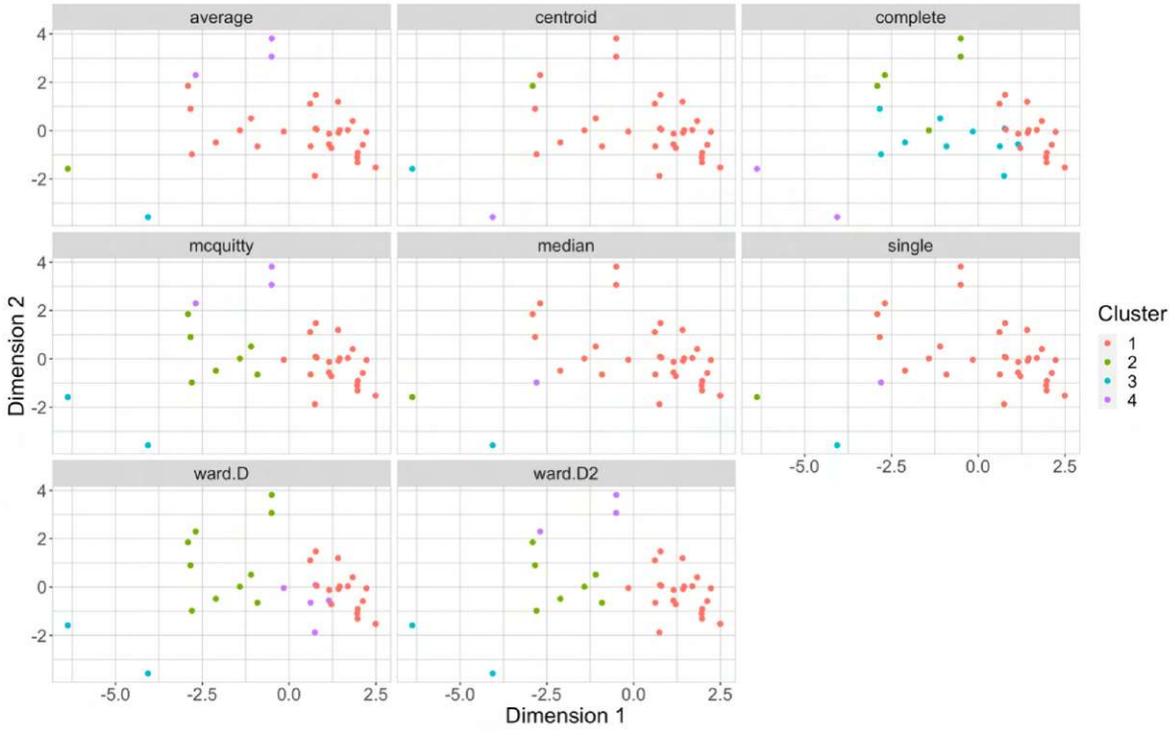


Figure 5.20: Patient assignment for various linkage methods using 4 clusters

5.2.2 Research Question 2.2

RQ 2.2: *What are the effects of using a different clustering method (e.g., fuzzy or robust methods)?*

While hierarchical clustering provides one option to cluster the patient cohort, other methods can be also used to get alternative solutions for this task. One widely used method is the *k-means* algorithm. In this approach, clusters are constructed by optimizing the position of a fixed number of cluster centers and assigning each observation to the closest cluster center. Similarly to the analysis performed in the previous section, we can use an elbow plot to get a first indication about the optimal number of clusters. As it can be seen in Figure 5.21, the elbow plot shows that up to 3 clusters, each additional cluster provides a substantial improvement. This is consistent with the observations made for hierarchical clustering as well.

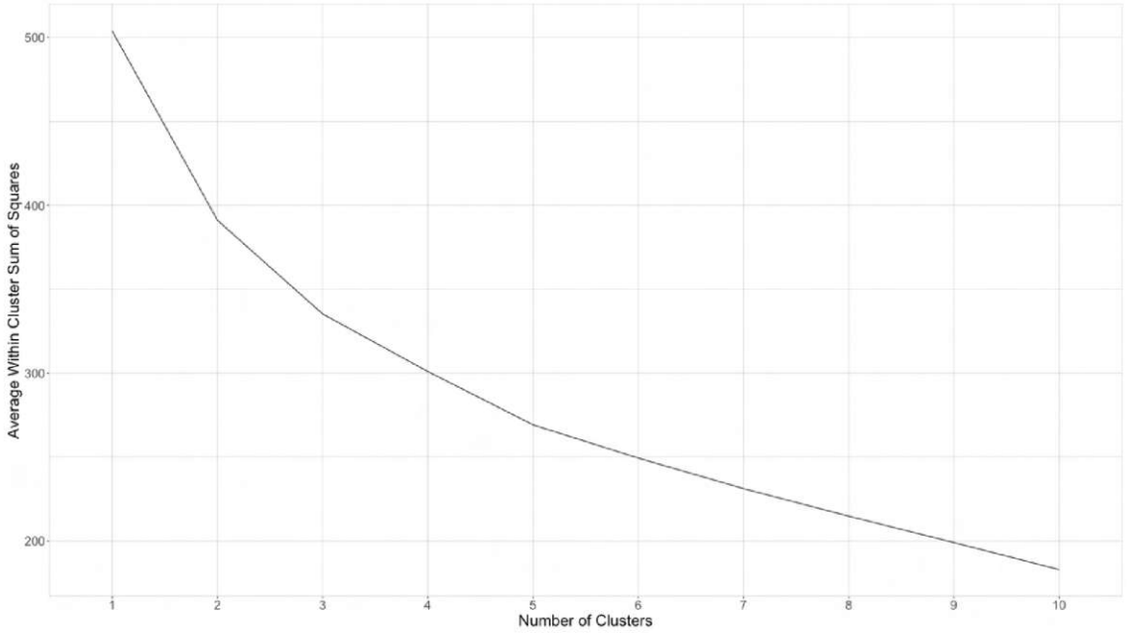


Figure 5.21: Elbow plot for the k-means clustering algorithm

To investigate the actual cluster assignments, we again use an abstraction of the distance matrix, which can be seen in Figure 5.22. Although the k-means algorithm uses the actual patient descriptors as input data rather than the distance matrix, this type of visualization enables a direct comparison with the clustering results for the hierarchical clustering. Moreover, additional symbols denote each cluster center optimized by the algorithm. It is also worth noting that this algorithm can often get stuck at local optima, depending on where the cluster centers were first initialized. Therefore, the clusters presented here are an aggregation of multiple iterations of the algorithm to ensure that the overall results represent the optimal solution. The cluster assignments are identical to those of hierarchical clustering under Euclidean distance with the linkage methods of *mcquitty* or *ward.d/ward.d2*. As mentioned earlier, cluster structures of this form are consistent with our goal of dividing the cohort into subgroups that are then used to make predictions.

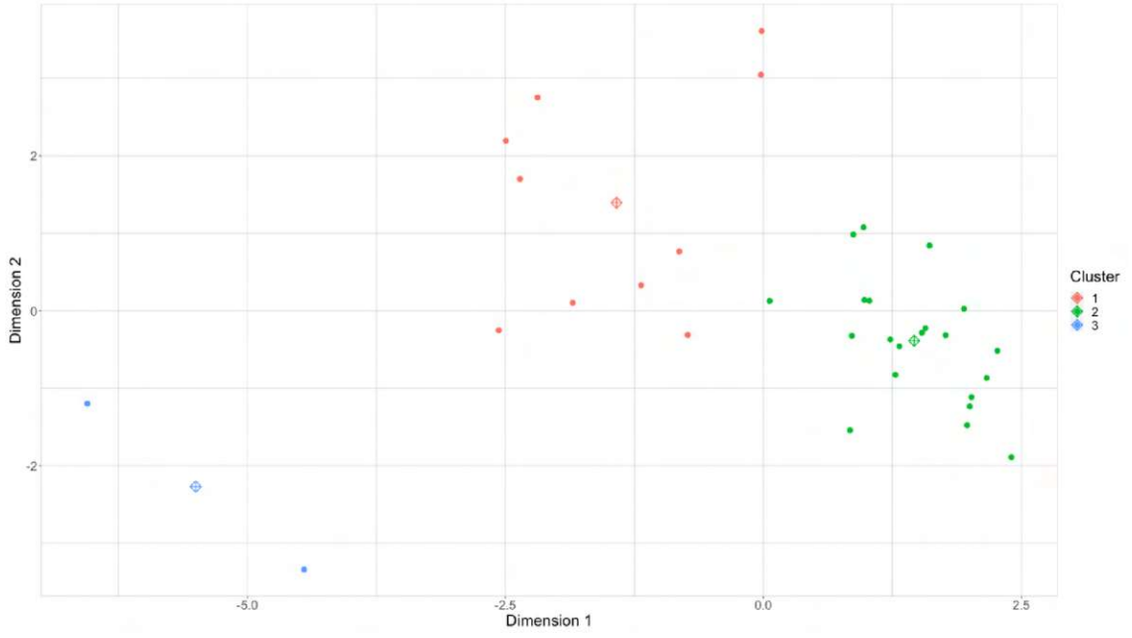


Figure 5.22: Cluster assignment made by k-means algorithm with 3 clusters

As an alternative to k-means, the k-medoids algorithm offers modifications that may be beneficial for our use case. The two main differences are that k-medoids always chooses one of the observations as the cluster centers and that it minimizes the sum of pairwise dissimilarities instead of the sum of squared Euclidean distances. This modification in particular makes it more robust to outliers, since their contribution is not further magnified. Similarly to k-means, we use an elbow plot to examine the amount of improvement by each additional cluster, as illustrated in Figure 5.23. In this case, a cluster number of 4 appears to be an optimal setting. However, to allow a comparison with the results for k-means, Figure 5.24 and Figure 5.25 show the results for both 3 and 4 clusters, respectively. When comparing the 3-cluster setting for k-means and k-medoids, significant differences in the results can be seen. In particular, the two observations identified as a separate cluster by k-means are part of a larger cluster in this case. Instead, four other patients are identified as a separate clustering by k-medoids. However, if we increase the number of clusters to 4, the two observations mentioned above are also grouped in a separate cluster. This has the effect of shifting the center of their former cluster, as highlighted by squares around the patients selected as the cluster center, and creating a slight deviation in its composition.

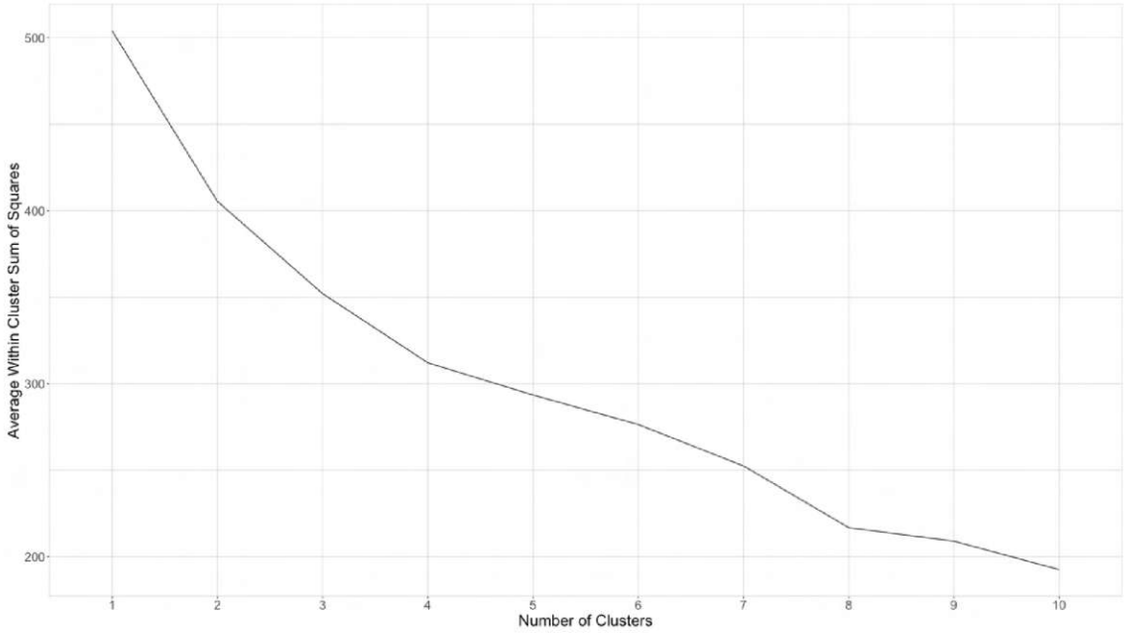


Figure 5.23: Elbow plot for the k-medoids clustering algorithm

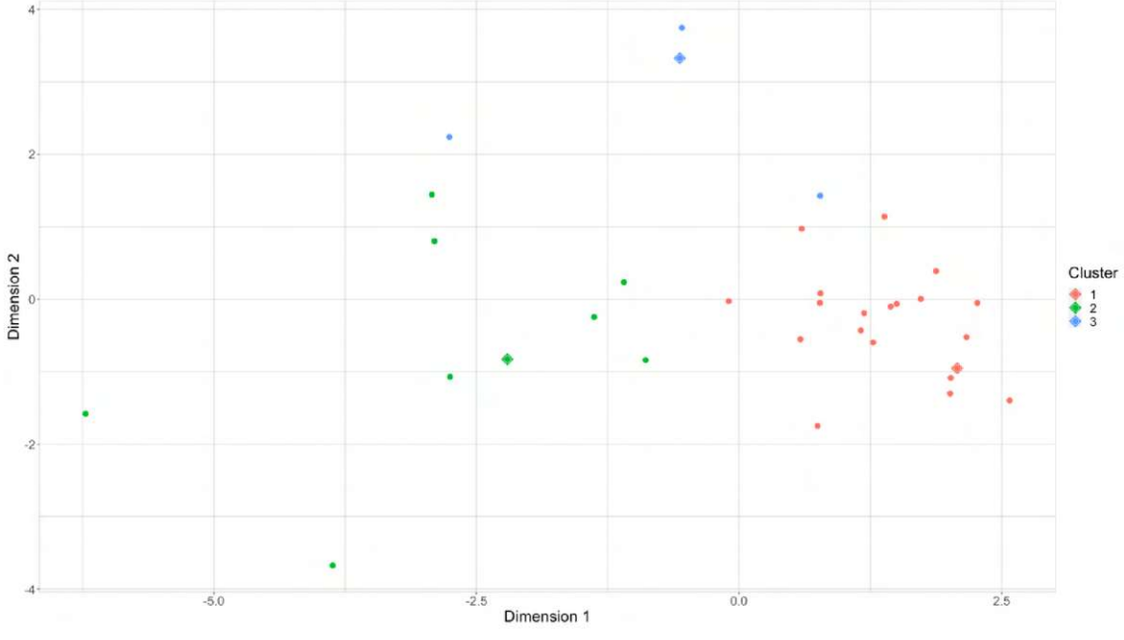


Figure 5.24: Cluster assignment made by k-medoids algorithm with 3 clusters

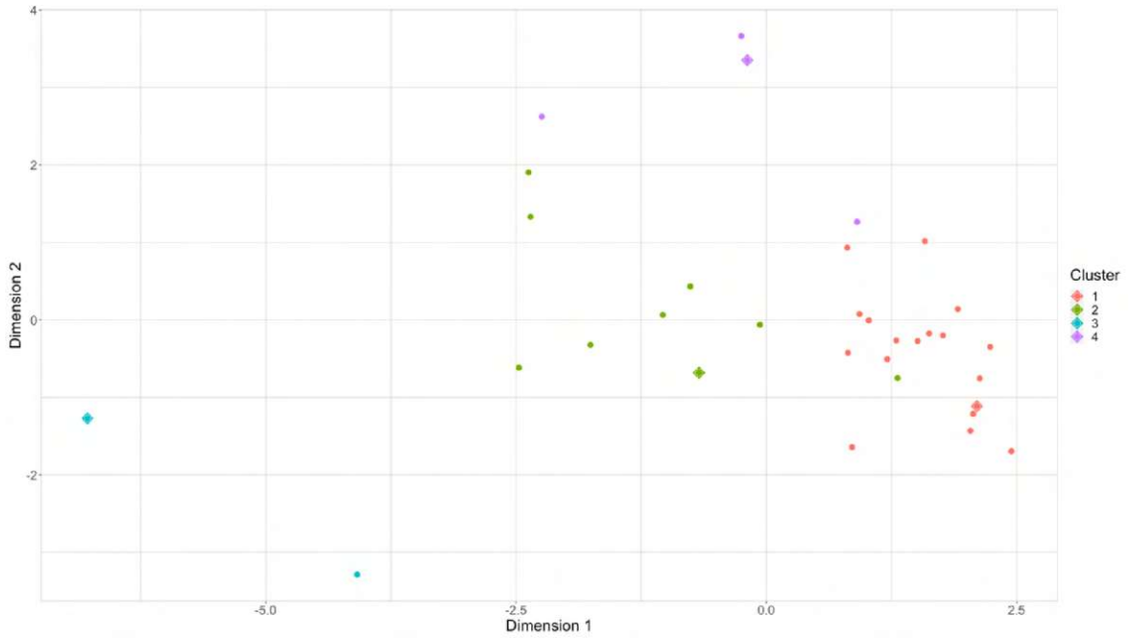
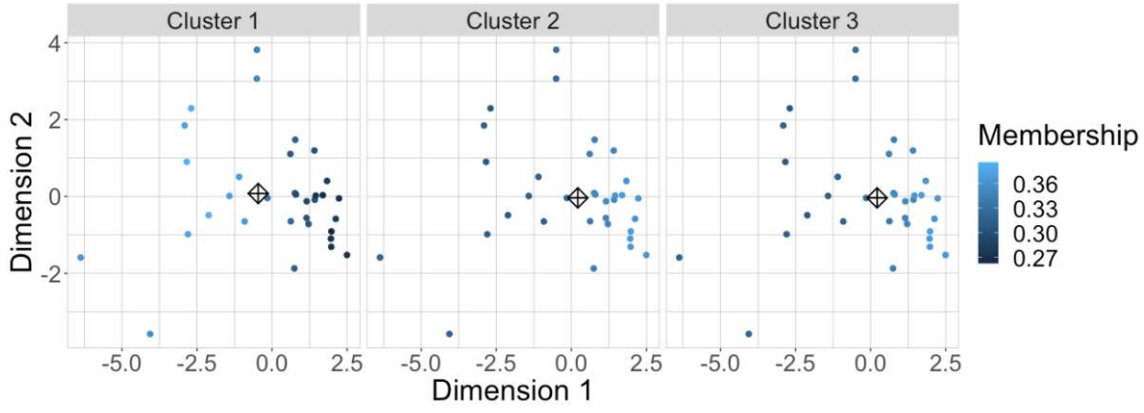
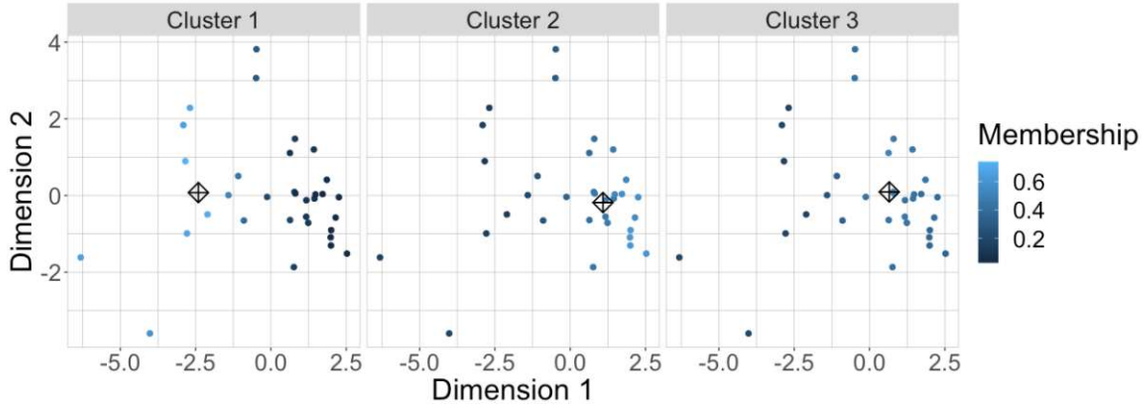


Figure 5.25: Cluster assignment made by k-medoids algorithm with 4 clusters

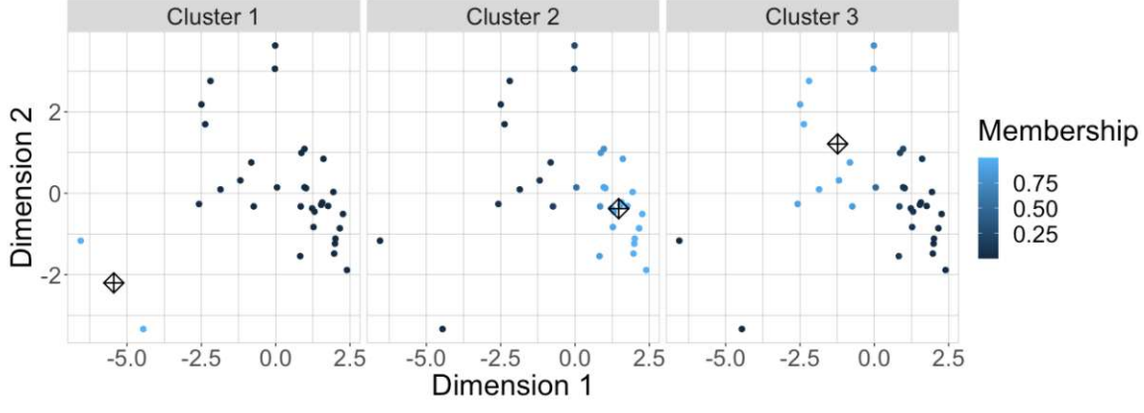
The above described methods assign each observation to a single cluster exclusively. However, there are alternative approaches that assign specific membership coefficients or probabilities of belonging to different clusters to each observation. Two such methods that will be examined in this section are *fuzzy k-means* and *model-based clustering*. Fuzzy k-means, as shown in Figure 5.26, involves a parameter m that determines the degree of fuzziness. As the value of m approaches 1, the algorithm converges to a binary assignment of 0 and 1 for each observation. This can be seen in Figure 5.26, where under the default setting of $m = 2$, the observations are assigned to each cluster with an almost equal degree. As m decreases, the separation between clusters increases, and the assignment approximates the results of the k-means algorithm. Theoretically, the prediction workflow presented in this work could be adapted to include all patients but with varying importance, for example by weighting their contribution to the generated shape variation samples. However, we can see that as the degree of fuzziness increases, it essentially equals using the entire cohort for the predictions, while a low level increasingly approximates a hard clustering. As an additional limitation, the fuzzy k-means algorithm has the disadvantage of being unstable, where re-running the algorithm may not result in identical membership coefficients. As a result, patients might be assigned to the clusters with a changing degree, which makes their interpretability less reliable as well.



(a) Fuzzy k-means membership coefficients with $m = 2$



(b) Fuzzy k-means membership coefficients with $m = 1.5$



(c) Fuzzy k-means membership coefficients with $m = 1.25$

Figure 5.26: Fuzzy k-means clustering assignment depending on the degree of fuzziness (m)

Another method that relies on assignment probabilities is model-based clustering. As discussed in Section 4.1.3, this method assumes that the observations in the dataset come

from different statistical distributions, representing the clusters it aims to identify. This property makes it also more likely to identify and separate outliers, as we can also see in Figure 5.27. Here, despite the probability based cluster assignment, the method resulted in three clusters without any estimated overlap and each observation was assigned 100% to its cluster. It is also noteworthy, that as we have argued earlier, such unbalanced clustering results are often not optimal for our prediction workflow, as certain patients have no additional observations available to use for the predictions, while others include almost the entire cohort.

Overall, we conclude for **RQ 2.2**, that the use of alternative clustering methods might provide advantages compared to hierarchical clustering, such as being less sensitive to potential outliers. Nonetheless, at the current stage of the patient cohort, even fundamentally different clustering methods tend to provide clusters with similar patterns.

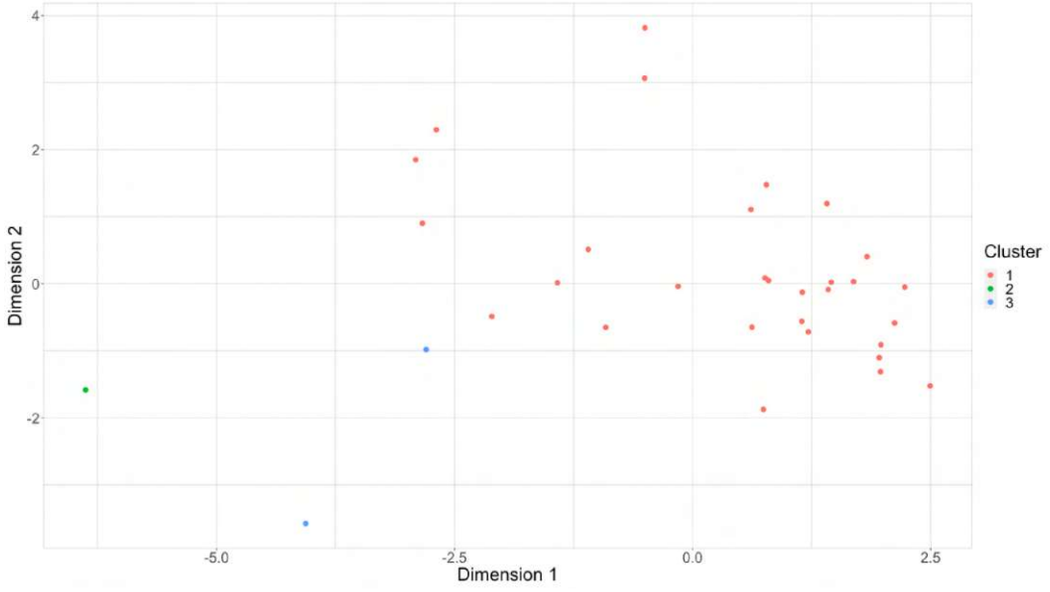


Figure 5.27: Model-based clustering result using three clusters

5.2.3 Joint Analysis of RQ 2.1 and RQ 2.2

While the previous two sections have addressed different techniques for performing clustering, they all share the common goal of dividing the cohort into subgroups that can then be used as input for the prediction workflow. Therefore, research questions **RQ 2.1** and **RQ 2.2** can be further analyzed jointly to investigate how different clustering results affect the prediction workflow presented in this paper. A first step in this analysis is to investigate how well different clustering approaches overlap in their results. So far, we have seen that the optimal number of clusters is generally 3 or 4, with slight variations in the cluster assignment patterns for the patients. To summarize these findings in a single visualization, we use a distance matrix representation that captures the similarity of the cluster assignment of different patients in the cohort. If two patients are always

clustered together, their distance is 0, and the patients are placed in the same location in the scatter plot. On the other hand, if two patients are never clustered together, their distance reaches a maximum of 1. All other cases are measured by an appropriate distance value in between.

To better understand this approach, Figure 5.28 shows the results for different number of clusters between 2 and 5. It can be seen that using only 2 clusters tends to separate some patients that we have already encountered frequently as possible outliers. The remaining observations tend to be clustered together, as indicated by the large number of overlapping points in the scatter plot. If we increase the number of clusters to 3, we not only separate the possible outliers, but also achieve a more or less consistent division of the remaining patients into two clusters. Here, the similarity of the cluster assignments across different settings seems to indicate three clearly separable groupings. As we increase the number of clusters to 4 or even 5, the similarity of patients in the scatter plot becomes more widespread, with greater variation in their cluster assignments. Overall, although we expect differences in the cluster assignments when using different settings, these deviations should be minimal if clusters are indeed present in the data and the number of these clusters is chosen appropriately. Thus, the results provide further evidence that this cohort of patients is best suited for clustering into 3 distinct patient groups. A detailed visualization for this setting is shown in Figure 5.29.

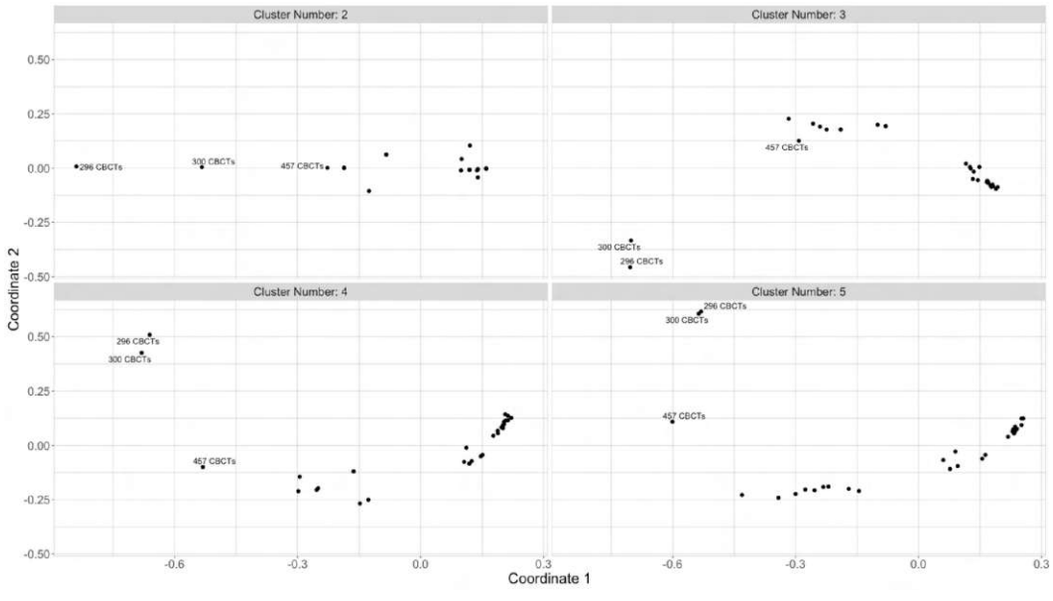


Figure 5.28: Cluster assignment similarity under different clustering settings using 2-5 clusters

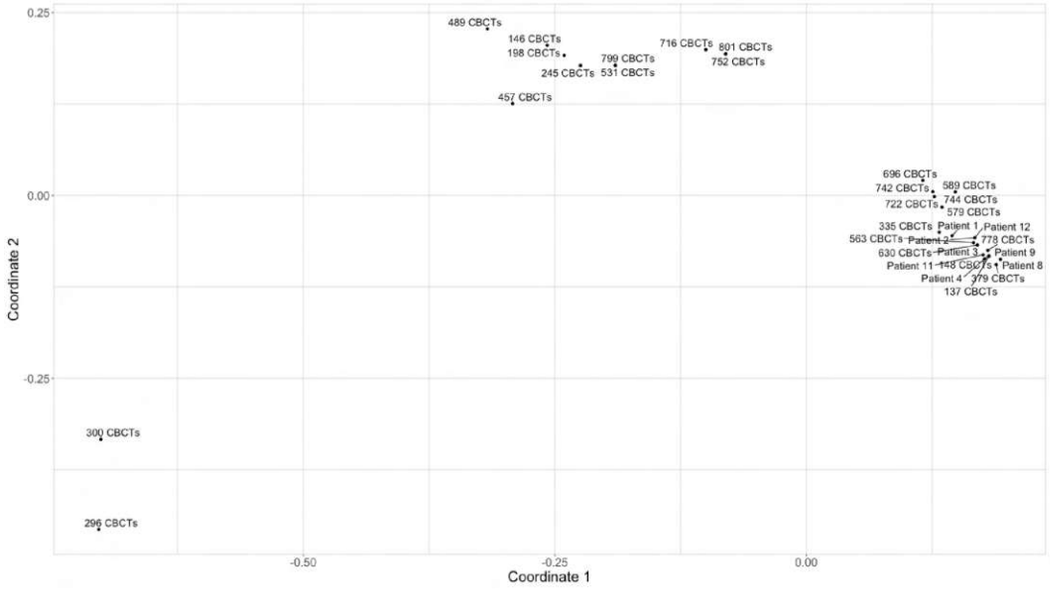


Figure 5.29: Cluster assignment similarity under different clustering settings using 3 clusters

Next, we focus on the impact of the different clustering results on the performance of the prediction workflow. As we have seen, dividing the cohort into three clusters leads to more or less consistent results with some variation in the groupings. Accordingly, we do not expect substantial deviations in prediction performance. However, it is of interest to see if there are any visible patterns in the data. For example, we have argued that an even split of the cohort is more desirable for our workflow. To do this, we first look at the standard deviation of the predictions under different settings. To exclude the influence of changes in the mean shape of the patient of interest, we limit the analysis to a 3 timesteps setting, which would be a reasonable real world scenario. We then calculate a standard deviation based on the prediction performances under different clustering approaches for each combination of patient, organ, and quantile of variation. Figure 5.30 summarizes the results for each quantile in a boxplot. In accordance with our expectations, the more variation we consider and add to the mean shape, the larger the deviations in the predictive performance. However, with the exception for the quantiles of 0 and 1, we can only observe small deviations that affect the performance by a few percent at most. Larger deviations are only visible for the two extreme cases of added variation. Therefore, we focus on these two cases in the further investigation.

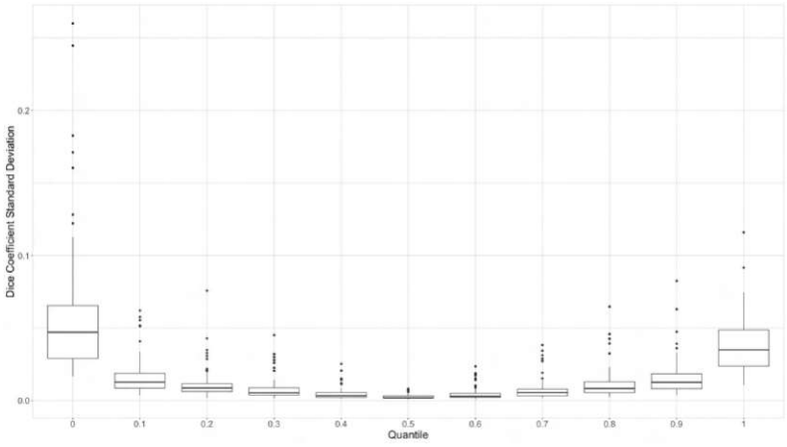


Figure 5.30: Standard deviation of the prediction performance using different clustering settings for each quantile of shape variation

Corresponding to the analysis presented for the hierarchical clustering, we now proceed to address the observations made there with respect to the prediction performance under different distance and linkage methods. Figure 5.31 and 5.32 summarize the performance achieved by the prediction workflow under different hierarchical clustering settings and aggregate them separately for different distance measures. While we originally concluded that all distance measures satisfy the majority of our requirements, with the exception of the maximum distance measure, the results do not indicate that any of the distance measures achieves a better performance overall. It is worth noting that for a quantile above 0.5, the mean shape part of the prediction always overlaps with the target, since we are only increasing it in size. On the other hand, when the quantile is below 0.5, especially in the extreme case of 0, it is possible that the prediction shrinks the mean shape to the point where it disappears and a Dice coefficient of 0 is returned.

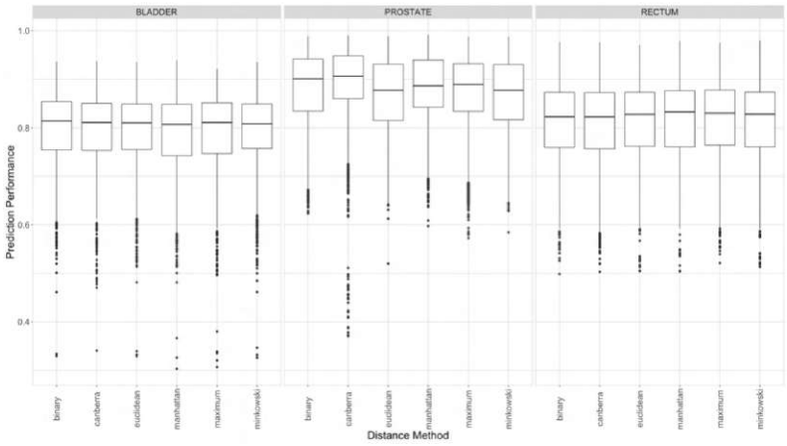


Figure 5.31: Prediction performance achieved by different distance methods for a variation quantile of 1

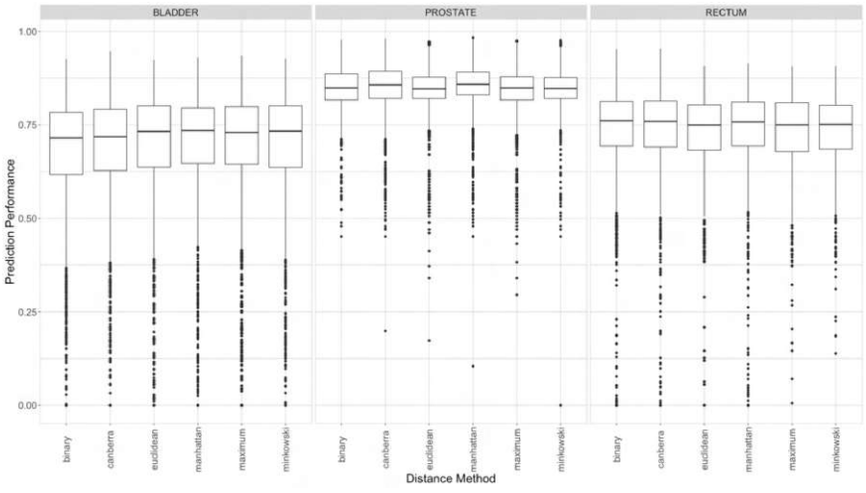


Figure 5.32: Prediction performance achieved by different distance methods for a variation quantile of 0

To investigate the importance of the linkage method, we next limit the analysis to the Euclidean distance, which corresponds to the analysis performed in Section 5.2.1. In Figure 5.33 and 5.34, the prediction performance achieved under this distance method is summarized in separate boxplots for different linkage methods. Here we can see that the linkage methods of *centroid*, *median*, and *single* perform slightly worse than the other alternatives. While this is in agreement with our assumption from Section 5.2.1, the difference is not large enough to draw general conclusions about the superiority of the other settings. In all the results presented here, it is noteworthy that the clustering is always optimized primarily for the bladder, since it is the dominant shape in the patient descriptor used as input.

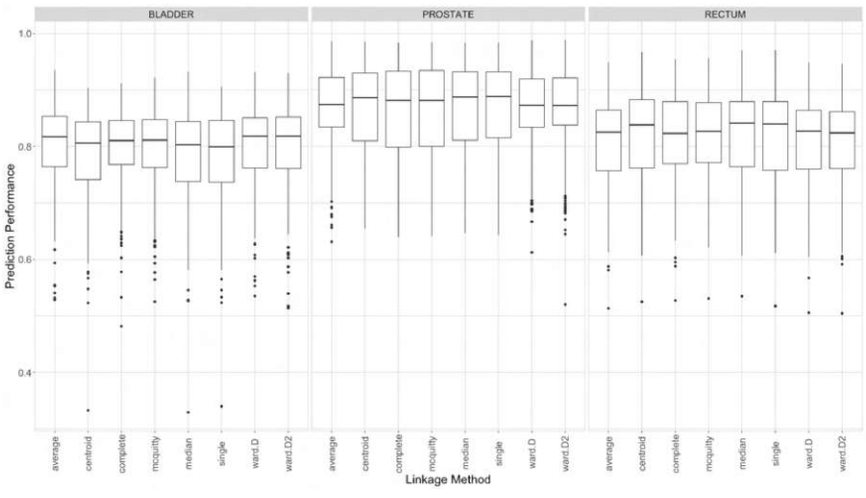


Figure 5.33: Prediction performance achieved by different linkage methods for a variation quantile of 1 (using Euclidean distance)

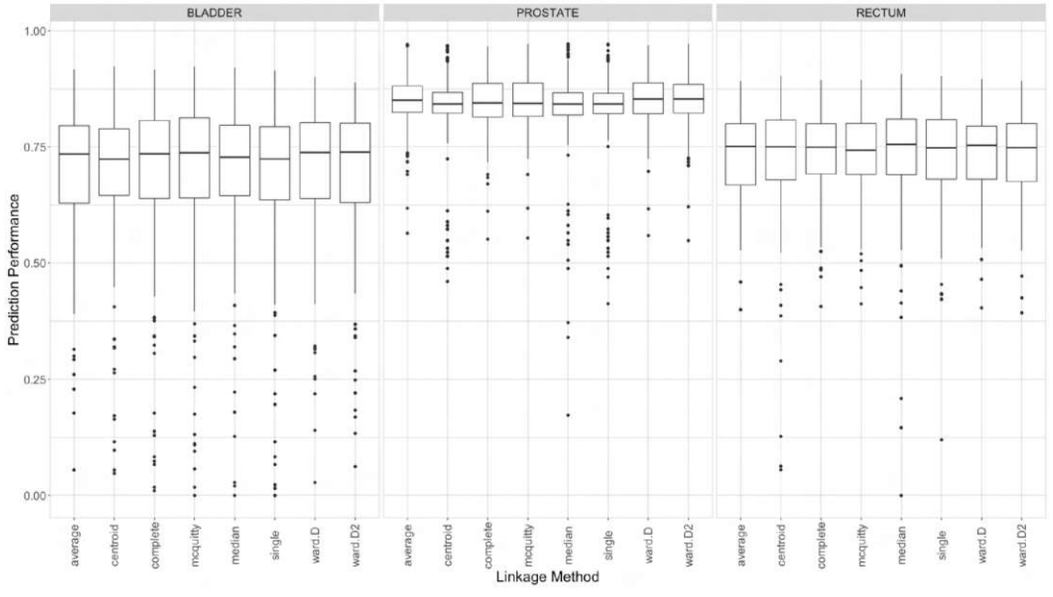


Figure 5.34: Prediction performance achieved by different linkage methods for a variation quantile of 0 (using Euclidean distance)

Summarizing our findings for **RQ 2.1** and **RQ 2.2**, we conclude that while the use of different clustering techniques can have an impact on the composition of the clusters, these changes have only a limited effect on the prediction performance itself. The largest variations in performance are observed in the prediction of the shape variability quantiles of 0 and 1. However, there is no single setting here that would consistently outperform all others.

5.2.4 Research Question 2.3

RQ 2.3: *How disruptive is the inclusion of a new observation with respect to existing clusters?*

As previously highlighted, for the exact implementation of the clustering approach, there are a variety of alternatives to choose from, many of which we have presented in Section 4.1.3 and further explored in **RQ 2.1** and **RQ 2.2**. These methods all share the common goal of identifying groups of similar patients, and achieve this by using different approaches that can lead to slightly different results. However, clustering methods in general are known to be sensitive to changes in the data, where even the inclusion or exclusion of individual observations can disrupt the previously identified groupings. In this research question, we simulate this problem by excluding individual observations and examining their effects on the cluster assignments. In answering this research question, we focus primarily on the settings that have previously been identified as optimal choices, but also aim for some general insights addressing all available alternatives.

The first important finding is that the number of clusters plays an important role in the stability of the clusters—something we have already seen in **RQ 2.1** and **RQ 2.2**. In the context of this research question, when focusing on hierarchical clustering under Euclidean distance and complete linkage, we can see that two clusters provide a clear separation of the cohort into two groups, as shown in Figure 5.35. More importantly, under these settings, the exclusion of any single patient from the cohort would not change the overall cluster assignment of patients.

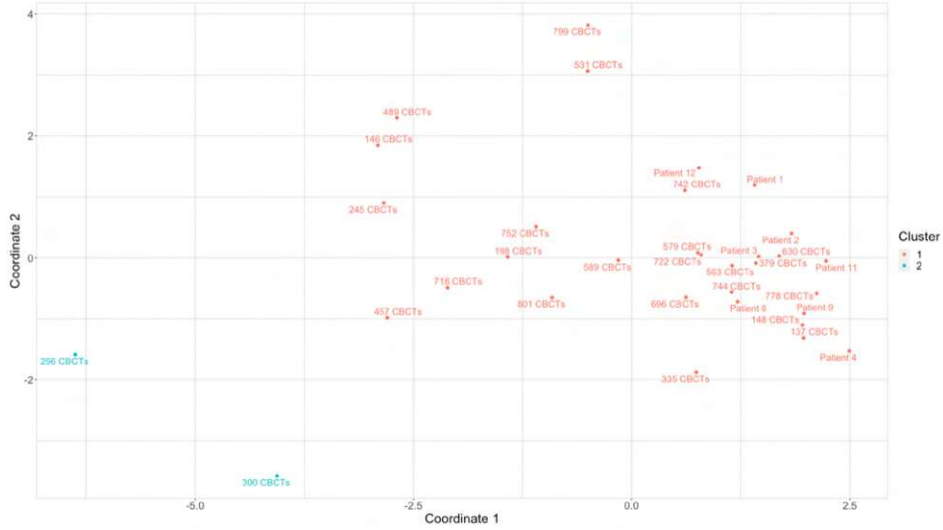


Figure 5.35: Patient cluster assignments using hierarchical clustering (Euclidean distance and complete linkage) and two clusters

However, we also know that the stability of cluster assignments is strongly influenced by possible outliers. One of the worst possible scenarios for this research question would be the inclusion or exclusion of a new patient who is significantly different from all other observations and thus represents a new, singular outlier in the cohort. In most of the methods investigated, we would expect such an observation to be isolated as its own cluster, resulting in a significant change in cluster assignments. In such cases, of course, the number of optimal clusters would have to be reevaluated. Using a two cluster setting, a similar case would be the result of excluding the two patients of *296 CBCTs* and *300 CBCTs*. Figure 5.36 shows the cluster assignments under these conditions, highlighting the disruption in the cluster assignments compared to Figure 5.35. However, since our analysis focuses primarily on the inclusion or exclusion of individual patients for this research question, we anticipate that there is no single outlier in our cohort that would systematically affect the analysis.

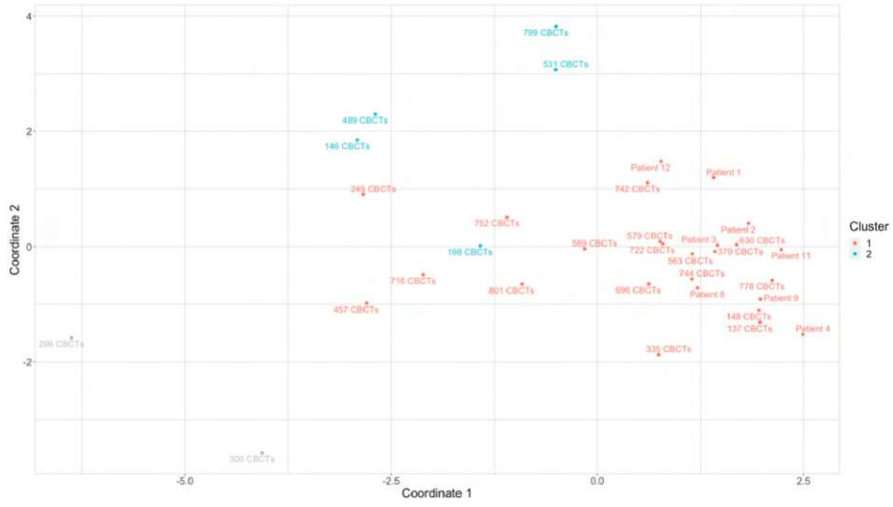


Figure 5.36: Patient cluster assignments using hierarchical clustering (Euclidean distance and complete linkage) and two clusters, with patients *296 CBCTs* and *300 CBCTs* excluded

In the next step, we focus on the setting with 3 clusters, which has proven to be an optimal choice for the patient cohort at its present stage. In this case, individual patients can influence the initially identified clusters in a more visible way. Figure 5.37 and 5.38 illustrate this with the exclusion of a selected patient using hierarchical clustering with Euclidean distance and complete linkage. Figure 5.37 shows the initial clusters based on the entire cohort, while Figure 5.38 shows the results when excluding Patient *457 CBCTs*. It is clear that an observation with the right characteristics can be crucial for the cluster assignments and thus lead to visible disruptions when included or excluded.

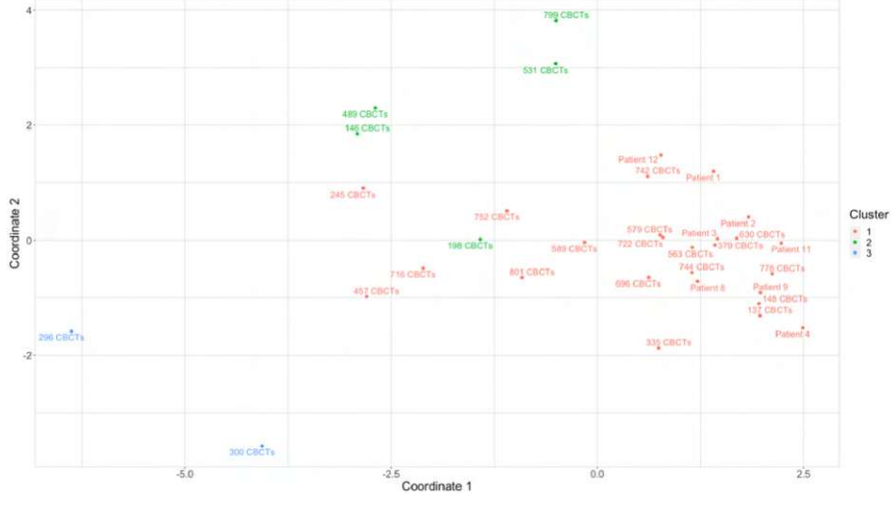


Figure 5.37: Patient cluster assignments using hierarchical clustering (Euclidean distance and complete linkage) and three clusters

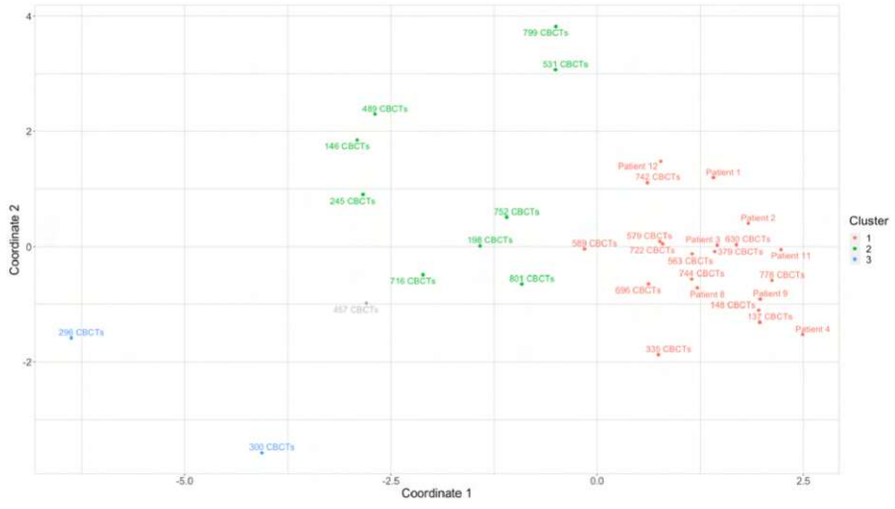


Figure 5.38: Patient cluster assignments using hierarchical clustering (Euclidean distance and complete linkage) and three clusters, with patient *457 CBCTs* excluded

It should be noted that the conclusions drawn from Figures above apply only to hierarchical clustering using Euclidean distance and complete linkage, settings we previously identified as preferable ones. From the analysis presented in **RQ 2.1**, we know that the linkage method in particular has a major impact on the importance assigned to individual patients. Based on our observations, certain settings, such as the complete linkage used above, prefer more balanced cluster sizes, while others, such as the centroid linkage, tend to separate individual patients as outliers. To provide insights into how these characteristics affect the results for this research question, we provide an overview of the observed cluster disruption under different clustering settings in Table 5.2. In this table, we summarize the number of changes in cluster assignments when excluding individual patients under different clustering settings. Furthermore, since the number of switches in assignment is also tightly related to the size of the clusters, we highlight the average size of the three clusters. Note, that the cluster numberings are independent across different clustering settings and should not be taken as a basis for comparison. The key insight in Table 5.2 can be found at the two ends of it. On the one hand, in the upper section we can observe settings that are less stable and can lead to larger variations in the cluster assignments. Ultimately, the larger number of variations also require more balanced cluster sizes that enable it. Furthermore, if we inspect the individual settings for these observations, we can see that they are ones, that by definition place an important role on individual patients or even individual positions in their shape descriptors during the calculations, leading to higher instability. On the other hand, at the bottom of the table we can find settings that are hardly affected by the inclusion or exclusion of individual patients from the cohort. All these are settings that tend to identify some patients as outliers and assign them as their own clusters. This is also visible in the average cluster sizes for these settings. This property makes them robust against changes in composition of the cohort, where in most cases only the inclusion or exclusion of an outlier itself has an effect on the cluster

assignments. From a qualitative point of view, neither of the two ends of the spectrum described above are desirable for our use case. As previously outlined in our analysis, a suitable choice of clustering settings could be the combination of Euclidean distance with complete linkage (see Section 5.2.1). In Table 5.2 we have highlighted this setting, which provided a reasonable stability at the middle of the spectrum in this analysis as well.

Overall, in **RQ 2.3** we have identified three key aspects, that affect the potential disruption caused by the inclusion or exclusion of individual patients in the cohort. First, the number of clusters, which can have a major impact on cluster stability. Second, the distribution of patients, especially the possibility of outliers in the cohort. And third, the correct choice of clustering parameterization. All these aspects and the use-case specific combination of them can have a large impact on the stability of cluster assignments of individual patients when additional observations are introduced to the cohort.

Algorithm	Link. Method	Dist. Method	Overlap	Cluster #1	Cluster #2	Cluster #3
kmeans	-	-	24.30	6.88	7.88	17.24
hierarchical	ward.D2	binary	25.55	11.58	14.21	6.21
hierarchical	mcquitty	canberra	27.45	14.67	8.70	8.64
hierarchical	ward.D	binary	28.55	12.06	9.09	10.85
hierarchical	mcquitty	binary	28.88	20.48	9.91	1.61
kmedoids	-	-	29.06	23.73	5.64	2.64
hierarchical	median	binary	29.67	29.30	1.36	1.33
hierarchical	ward.D2	canberra	29.70	11.39	10.00	10.61
hierarchical	ward.D	canberra	29.85	11.64	9.58	10.79
hierarchical	complete	canberra	29.97	8.36	15.82	7.82
hierarchical	average	canberra	30.18	12.48	16.67	2.85
hierarchical	complete	maximum	30.42	20.91	9.09	2.00
hierarchical	average	binary	30.45	5.27	24.33	2.39
hierarchical	mcquitty	euclidean	30.48	21.88	8.18	1.94
hierarchical	mcquitty	minkowski	30.48	21.88	8.18	1.94
hierarchical	complete	binary	30.58	9.73	19.18	3.09
hierarchical	average	euclidean	30.76	27.12	1.76	3.12
hierarchical	average	minkowski	30.76	27.12	1.76	3.12
hierarchical	complete	manhattan	31.03	28.15	2.18	1.67
hierarchical	median	maximum	31.03	30.00	1.00	1.00
hierarchical	mcquitty	maximum	31.06	23.30	7.70	1.00
hierarchical	**complete**	**euclidean**	**31.12**	**24.39**	**5.67**	**1.94**
hierarchical	complete	minkowski	31.12	24.39	5.67	1.94
hierarchical	ward.D	maximum	31.15	7.27	10.70	14.03
hierarchical	single	canberra	31.24	28.82	1.94	1.24
hierarchical	mcquitty	manhattan	31.27	29.67	1.33	1.00
hierarchical	average	maximum	31.30	22.00	9.00	1.00
hierarchical	ward.D2	maximum	31.39	8.67	9.91	13.42
hierarchical	median	canberra	31.58	30.00	1.00	1.00
hierarchical	ward.D	euclidean	31.76	20.48	9.52	2.00
hierarchical	ward.D	minkowski	31.76	20.48	9.52	2.00
hierarchical	ward.D2	euclidean	31.79	20.39	9.61	2.00
hierarchical	ward.D2	minkowski	31.79	20.39	9.61	2.00
hierarchical	ward.D	manhattan	31.82	20.42	9.58	2.00
hierarchical	ward.D2	manhattan	31.82	20.42	9.58	2.00
hierarchical	single	maximum	31.88	30.00	1.00	1.00
hierarchical	average	manhattan	31.94	30.00	1.00	1.00
hierarchical	centroid	binary	31.94	30.00	1.00	1.00
hierarchical	centroid	canberra	31.94	30.00	1.00	1.00
hierarchical	centroid	euclidean	31.94	30.00	1.00	1.00
hierarchical	centroid	manhattan	31.94	30.00	1.00	1.00
hierarchical	centroid	maximum	31.94	30.00	1.00	1.00
hierarchical	centroid	minkowski	31.94	30.00	1.00	1.00
hierarchical	median	euclidean	31.94	30.00	1.00	1.00
hierarchical	median	manhattan	31.94	30.00	1.00	1.00
hierarchical	median	minkowski	31.94	30.00	1.00	1.00
hierarchical	single	euclidean	31.94	30.00	1.00	1.00
hierarchical	single	manhattan	31.94	30.00	1.00	1.00
hierarchical	single	minkowski	31.94	30.00	1.00	1.00
hierarchical	single	binary	31.97	29.06	1.94	1.00
modelBased	-	-	31.97	29.09	1.00	1.91

Table 5.2: Disruption in cluster assignments caused by patients excluded from the cohort (numeric columns measuring the average overlap of the cluster assignments and the average cluster sizes)

CHAPTER 6

Conclusion

This thesis has focused on examining and improving the prediction workflow first presented by Furmanová et al. [FMCM+21], where a pelvic organ shape variability prediction approach, involving the clustering of the patient cohort has been proposed. Throughout our investigations, we have identified several areas in which the original workflow can be enhanced, as well as confirmed the effectiveness of certain elements of the original design. Furthermore, we have explored the effects of potential real-world issues on the prediction workflow. In addition, we have developed a visual analytics interface to facilitate the examination of the clustering and prediction workflow. In this chapter, we will provide a comprehensive overview of the work that has been conducted, including the results of our analyses and potential questions of interest for future research.

6.1 Research Findings and Contribution

Building on the list of research questions defined in Section 1.2, we first summarize our findings with respect to area of shape descriptors and then the topic of clustering within the prediction workflow presented in this work. Regarding the shape descriptors used to represent the individual patient's pelvic organs, we compared the methods used in PREVIS with an alternative approach based on reducing the resolution of the voxels (see **RQ 1.1**). We found that the alternative method presented in this work provides higher quality shape descriptors with increased stability across patients and organs, both in terms of representation of the underlying organs as well as their reconstruction capability. Regardless of the method used, lower performance has been observed for the rectum, due to its more irregular and non-spherical shape. In subsequent research questions, we quantified the effects of missing slices on the quality of shape descriptors (see **RQ 1.2**). Our results showed that our method is robust for a low number of missing layers. However, missing layers in the core of the organs can have a significant impact on parts of our workflow and affect the upsampling workflow is used for a reconstruction of

the organ. Finally, we examined the impact of using different numbers of CT slices and quantile settings for the prediction of the organ shape variability of individual patients (see **RQ 1.3**). Here, we we have seen, that outliers in the CT scan series can have significant influences on he calculated mean shape of the organs and thus have a large effect on the prediction workflow as well. Furthermore we have seen, that the variation quantiles of 0 and 1, describing the most extreme cases of variation for individual patients, are the most challenging to accurately predict.

In the second part, we have focused on the topic of clustering. We compared different clustering methods and settings, both in terms of the resulting cluster assignment of the patients as well as the predictive performance they enable (see **RQ 2.1 and 2.2**). With respect to hierarchical clustering, we have identified distance measures that tend to more accurately represent the actual overlap between the organs compared. Similarly, we have identified differences in various linkage methods and their impact on the composition of the resulting clusters. In particular, we noted differences in the size of the clusters they identified and the way they were affected by outliers in the patient cohort. Overall, we concluded that the choice of Euclidean distance in combination with complete linkage—as used in the original publication by Furmanová et al. [FMCM+21]—provides clusters with optimal properties for our use case. Although we found visible differences in the cluster assignment of patients under different settings, these had only a limited effect on the predictive performance of the workflow. The largest deviations among the predictions were observed when the workflow was used to predict quantiles of 0 and 1, capturing the most extreme variations in the organs. Finally, we shifted our focus to the effects and disruptions that the inclusion or exclusion of individual patients might have on the previously identified clusters (see **RQ 2.3**). In this case, we found that individual patients may cause visible differences in cluster assignments, with the magnitude of the changes also depending on the specific clustering settings used. However at the current stage of the patient cohort, the cluster assignment changes generally only affect a limited number of patients, with the optimal number clusters remaining the same.

In addition to the quantitative analysis, we also provide a visual analytics dashboard, that supports the interactive exploration of the aforementioned aspects. Based on a more patient-oriented context, this application facilitated the qualitative exploration of the research questions with respect to individual patients. In particular, the visual representation of the clusters proved to be useful in identifying potential outliers and comparing cluster sizes. Therefore, such visualization dashboards could be used in the real world to find optimal settings for incoming new patients.

6.2 Limitations and Future Research

When interpreting the findings of this work, it is important to note certain limitations. One limitation is the small size of the patient cohort, which can affect the interpretation of the clustering results and the performance of the prediction workflow. In the long term, the augmentation of the patient cohort would likely lead to a more distinct

differentiation between clusters and a greater range of prediction performance produced by different settings. Another limitation is that the clustering approach used in this study is based solely on the shape descriptors and does not take into account other patient specific information that might be relevant. Future studies could explore the inclusion of demographic factors, such as age, as additional features in the clustering process. In particular with respect to the prostate, we expect to see a relationship between prostate size and age [ZQZ+13], which could improve the quality of the clusters. Additionally, the shape descriptors used in this work are centered according to the prostate, which limits the information about the prostate itself. Alternative centering methods, for example ones based on the surrounding bone structure, may preserve more information about the prostate shape and position variability and provide additional insights. Finally, the current study does not distinguish between organ movement, and organ size or shape variation as sources of variability. Future research could investigate a decomposition of these two and explore their individual degree of contribution.

6.3 Summary

In summary, our work has conducted a comprehensive examination of the prediction workflow first presented by Furmanová et al. [FMCM+21]. Our evaluations identified potential limitations and issues and made improvements to the initial implementation of the workflow. By conducting such thorough evaluations, we have addressed one of many prerequisites before considering the practical application of prediction workflows, such as PREVIS, in real-world analytical environments. The results of this work were also partially submitted and presented at EG VCBM 2022 under the title "Understanding the impact of statistical and machine learning choices on predictive models for radiotherapy" [BFR22], where the manuscript was awarded the best short paper award.

List of Figures

List of Tables

Bibliography

[ASYS06] Ceyhun Burak Akgül, Bülent Sankur, Yücel Yemez, and Francis Schmitt. A framework for histogram-induced 3d descriptors. In *2006 14th European Signal Processing Conference*, pages 1–5. IEEE, 2006.

[BBP10] Stef Busking, Charl P Botha, and Frits H Post. Dynamic multi-view exploration of shape spaces. In *Computer Graphics Forum*, volume 29, pages 973–982. Wiley Online Library, 2010.

[BCHR00] David J Brenner, Rochelle E Curtis, Eric J Hall, and Elaine Ron. Second malignancies in prostate carcinoma patients after radiotherapy compared with surgery. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 88(2):398–406, 2000.

[BEF84] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203, 1984.

[BFR22] Ádám Böröndy, Katarína Furmanová, and Renata Georgia Raidou. Understanding the impact of statistical and machine learning choices on predictive models for radiotherapy. 2022.

[BKS+11] E Budiarto, M Keijzer, PR Storchi, MS Hoogeman, L Bondar, TF Mutanga, HCJ de Boer, and AW Heemink. A population-based model to describe geometrical uncertainties in radiotherapy: applied to prostate cases. *Physics in Medicine & Biology*, 56(4):1045, 2011.

[BM13] Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385, 2013.

[CMMH+17] Oscar Casares-Magaz, Vitali Moiseenko, Austin Hopper, Niclas Johan Pettersson, Maria Thor, Rick Knopp, Joseph O Deasy, Ludvig Paul Muren, and John Einck. Associations between volume changes and spatial dose metrics for the urinary bladder during local versus pelvic irradiation for prostate cancer. *Acta Oncologica*, 56(6):884–890, 2017.

[CSB15] David RH Christie, Christopher F Sharpley, and Vicki Bitsika. Why do patients regret their prostate cancer treatment? a systematic review of regret after treatment for localized prostate cancer. *Psycho-Oncology*, 24(9):1002–1011, 2015.

[dCBP+18] Renaud de Crevoisier, Mohamed Amine Bayar, Pascal Pommier, Xavier Muracciole, Françoise Pêne, Philippe Dudouet, Igor Latorzeff, Véronique Beckendorf, Jean-Marc Bachaud, Agnès Laplanche, et al. Daily versus weekly prostate cancer image guided radiation therapy: phase 3 multicenter randomized trial. *International Journal of Radiation Oncology* Biology* Physics*, 102(5):1420–1429, 2018.

[Dic45] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[FGM+20] Katarína Furmanová, Nicolas Grossmann, Ludvig P Muren, Oscar Casares-Magaz, Vitali Moiseenko, John P Einck, M Eduard Gröller, and Renata G Raidou. Vapor: Visual analytics for the exploration of pelvic organ variability in radiotherapy. *Computers & Graphics*, 91:25–38, 2020.

[FMCM+21] Katarína Furmanová, Ludvig P Muren, Oscar Casares-Magaz, Vitali Moiseenko, John P Einck, Sara Pilskog, and Renata G Raidou. Previs: predictive visual analytics of anatomical variability for radiotherapy decision support. *Computers & Graphics*, 2021.

[FR02] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[GCMM+19] Nicolas Grossmann, Oscar Casares-Magaz, Ludvig Paul Muren, Vitali Moiseenko, John P Einck, M Eduard Gröller, and Renata Georgia Raidou. Pelvis runner: Visualizing pelvic organ variability in a cohort of radiotherapy patients. In *VCBM*, pages 69–78, 2019.

[GTF+11] Suki Gill, Jessica Thomas, Chris Fox, Tomas Kron, Aldo Rolfo, Mary Leahy, Sarat Chander, Scott Williams, Keen Hun Tai, Gillian M Duchesne, et al. Acute toxicity in prostate cancer patients treated with and without image-guided radiotherapy. *Radiation oncology*, 6(1):1–7, 2011.

[HGP99] Patrick Hoffman, Georges Grinstein, and David Pinkney. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, pages 9–16, 1999.

[Hil35] David Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. In *Dritter Band: Analysis · Grundlagen der Mathematik · Physik Verschiedenes*, pages 1–2. Springer, 1935.

[HMSW04] Wolfgang Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and semiparametric models*, volume 1. Springer, 2004.

[HW79] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[IAR20] IARC. Cancer toady. `https://gco.iarc.fr/today/home`, 2020. Accessed: 2021-12-30.

[ID09] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates. *Human-Machine Interactive Systems*, pages 199–233, 2009.

[KCM+13] Mira Keyes, Juanita Crook, Gerard Morton, Eric Vigneault, Nawaid Usmani, and W James Morris. Treatment options for localized prostate cancer. *Canadian Family Physician*, 59(12):1269–1274, 2013.

[KFR03] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003.

[KLR+13] Paul Klemm, Kai Lawonn, Marko Rak, Bernhard Preim, Klaus D Tönnies, Katrin Hegenscheid, Henry Völzke, and Steffen Oeltze. Visualization and analysis of lumbar spine canal variability in cohort study data. In *VMV*, pages 121–128, 2013.

[KOJL+14] Paul Klemm, Steffen Oeltze-Jafra, Kai Lawonn, Katrin Hegenscheid, Henry Völzke, and Bernhard Preim. Interactive visual analysis of image-centric cohort study data. *IEEE transactions on visualization and computer graphics*, 20(12):1673–1682, 2014.

[KR90a] Leonard Kaufman and Peter J Rousseeuw. Finding groups in data: An introduction to cluster analysis–john wiley & sons. *Inc., New York*, 1990.

[KR90b] Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125, 1990.

[LLSE11] Sabine Landau, Morven Leese, Daniel Stahl, and Brian S Everitt. *Cluster analysis*. John Wiley & Sons, 2011.

[MD14] Neil E Martin and Anthony V D'Amico. Progress and controversies: Radiation therapy for prostate cancer. *CA: a cancer journal for clinicians*, 64(6):389–407, 2014.

[MLK+07] Vitali Moiseenko, Mitchell Liu, Sarah Kristensen, Gerald Gelowitz, and Eric Berthelet. Effect of bladder filling on doses to prostate and organs at risk: a treatment planning study. *Journal of Applied Clinical Medical Physics*, 8(1):55–68, 2007.

[PI+97] Markus Peura, Jukka Iivarinen, et al. Efficiency of simple shape descriptors. In *Proceedings of the third international workshop on visual form*, volume 5, pages 443–451, 1997.

[PLL99] José M Pena, Jose Antonio Lozano, and Pedro Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, 20(10):1027–1040, 1999.

[PP02] Tom Pickles and Norm Phillips. The risk of second malignancy in men with prostate cancer treated with or without radiation in british columbia, 1984–2000. *Radiotherapy and oncology*, 65(3):145–151, 2002.

[Raw19] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2):63, 2019.

[RBGR18] Oliver Reiter, Marcel Breeuwer, M Eduard Gröller, and Renata Georgia Raidou. Comparative visual analysis of pelvic organ segmentations. In *EuroVis (Short Papers)*, pages 37–41, 2018.

[RCMA+18] Renata G Raidou, Oscar Casares-Magaz, Artem Amirkhanov, Vitali Moiseenko, Ludvig P Muren, John P Einck, Anna Vilanova, and M Eduard Gröller. Bladder runner: Visual analytics for the exploration of rt-induced bladder toxicity in a cohort study. In *Computer Graphics Forum*, volume 37, pages 205–216. Wiley Online Library, 2018.

[RDCO+17] Richard Rios, Renaud De Crevoisier, Juan D Ospina, Frederic Commandeur, Caroline Lafond, Antoine Simon, Pascal Haigron, Jairo Espinosa, and Oscar Acosta. Population model of bladder motion and deformation based on dominant eigenmodes and mixed-effects models in prostate cancer radiotherapy. *Medical image analysis*, 38:133–149, 2017.

[RE19] Trevor J Royce and Jason A Efstathiou. Proton therapy for prostate cancer: A review of the rationale, evidence, and current state. In *Urologic Oncology: Seminars and Original Investigations*, volume 37, pages 628–636. Elsevier, 2019.

[SDPHM20] Leif-Erik D Schumacher, Alan Dal Pra, Sarah E Hoffe, and Eric A Mellon. Toxicity reduction required for mri-guided radiotherapy to be cost-effective in the treatment of localized prostate cancer. *The British journal of radiology*, 93(1114):20200028, 2020.

[SLHC14] Lotte Sander, Niels Christian Langkilde, Mats Holmberg, and Jesper Carl. Mri target delineation may reduce long-term toxicity after prostate radiotherapy. *Acta Oncologica*, 53(6):809–814, 2014.

[Sor48] Th A Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34, 1948.

[SRM+19] Matthias Schlachter, Renata G Raidou, Ludvig P Muren, Bernhard Preim, Paul Martin Putora, and Katja Bühler. State-of-the-art report: Visual computing in radiation therapy planning. In *Computer Graphics Forum*, volume 38, pages 753–779. Wiley Online Library, 2019.

[Tho53] Robert L Thorndike. Who belongs in the family. In *Psychometrika*. Citeseer, 1953.

[TWH01] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[VLBK+13] Tatiana Von Landesberger, Sebastian Bremm, Matthias Kirschner, Stefan Wesarg, and Arjan Kuijper. Visual analytics for model-based medical image segmentation: Opportunities and challenges. *Expert Systems with Applications*, 40(12):4934–4943, 2013.

[YB16] Slav Yartsev and Glenn Bauman. Target margins in radiotherapy of prostate cancer. *The British journal of radiology*, 89(1067):20160312, 2016.

[ZCM+99] Michael J Zelefsky, Diane Crean, Gig S Mageras, Olga Lyass, Laura Happersett, C Clifton Ling, Steven A Leibel, Zvi Fuks, Sarah Bull, Hanne M Kooy, et al. Quantification and predictors of prostate position variability in 50 patients evaluated with multiple ct scans during conformal radiotherapy. *Radiotherapy and oncology*, 50(2):225–234, 1999.

[ZLH+08] Michael J Zelefsky, Emily J Levin, Margie Hunt, Yoshiya Yamada, Alison M Shippy, Andrew Jackson, and Howard I Amols. Incidence of late rectal and urinary toxicities after three-dimensional conformal radiotherapy and intensity-modulated radiotherapy for localized prostate cancer. *International Journal of Radiation Oncology* Biology* Physics*, 70(4):1124–1129, 2008.

[ZQZ+13] Shi-Jun Zhang, Hai-Ning Qian, Yan Zhao, Kai Sun, Hui-Qing Wang, Guo-Qing Liang, Feng-Hua Li, and Zheng Li. Relationship between age and prostate size. *Asian journal of andrology*, 15(1):116, 2013.