Eurographics Workshop on Visual Computing for Biology and Medicine (2022)
T. Schultz and H.-Y. Wu (Editors)

*Short Paper*

# Understanding the impact of statistical and machine learning choices on predictive models for radiotherapy

Ádám Böröndy[1], Katarína Furmanová[2], and Renata Georgia Raidou[1]

[1]TU Wien, Austria, [2] Masaryk University, Czech Republic

**Abstract**

*During radiotherapy (RT) planning, an accurate description of the location and shape of the pelvic organs is a critical factor for the successful treatment of the patient. Yet, during treatment, the pelvis anatomy may differ significantly from the planning phase. A series of recent publications, such as PREVIS [FMCM*21], have examined alternative approaches to analyzing and predicting pelvic organ variability of individual patients. These approaches are based on a combination of several statistical and machine learning methods, which have not been thoroughly and quantitatively evaluated within the scope of pelvic anatomical variability. Several of their design decisions could have an impact on the outcome of the predictive model. The goal of this work is to assess the impact of alternative choices, focusing mainly on the two key-aspects of shape description and clustering, to generate better predictions for new patients. The results of our assessment indicate that resolution-based descriptors provide more accurate and reliable organ representations than state-of-the-art approaches, while different clustering settings (distance metric and linkage) yield only slightly different clusters. Different clustering methods are able to provide comparable results, although when more shape variability is considered their results start to deviate. These results are valuable for understanding the impact of statistical and machine learning choices on the outcomes of predictive models for anatomical variability.*

**CCS Concepts**

• ***Human-centered computing*** → ***Visual Analytics;*** • ***Applied computing*** → ***Life and medical sciences;***

## 1. Introduction

In radiotherapy (RT) planning, an accurate description of the location and shape of the pelvic organs is a critical factor for successful tumor treatment [SRM*19]. The use of sub-optimal settings might overexpose healthy organs to radiation—thus, to reduce the possibility of side effects, precise targeting of the cancerous area is sought. 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 RT 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. Furthermore, the extent of these changes tends to vary across patients.

A series of recent publications have examined alternative approaches for analyzing and predicting the pelvic organ variability of individual patients [RCMA*18, FGM*20, FMCM*21]. One of these approaches, *PREVIS*, proposed by Furmanová et al. [FMCM*21] uses a set of cancer patients from a retrospective cohort with known variability to generate personalized predictions for new patients (Figure 1). The approach is based on a combination of several statistical and machine learning methods, which have not yet been thoroughly and quantitatively evaluated within the scope of pelvic anatomical variability. Several of the design decisions of *PREVIS* could have an impact on the outcome of the predictive model and, thus, on RT decision making.

The contribution of this work is an assessment approach to understand and improve the robustness of predictive models of the anatomical variability of patients within the course of RT. In essence, our work aims at answering: *What is the impact of alternative choices, employed for the prediction of anatomical variability, on the final outcome of the predictive workflow of PREVIS?* We focus specifically on the following two key-aspects (Figure 1):

**Expression of organ shapes by 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. The organ segmentations in the CT scans are converted into mathematical shape descriptors that capture the presence of an organ at specific positions. It is anticipated that the choice of shape descriptor has a significant influence for the rest of the workflow. Practically, we are interested in understanding *what the effects of using different shape descriptors are*. 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 organ and how accurate predictions it enables.

**Clustering of past patients:** The prediction model is based on identifying clusters of past patients with similar organ variability

DOI: 10.2312/vcbm.20221188

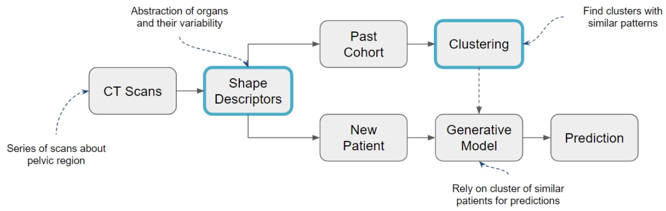

Abstraction of organs and their variability
CT Scans
Shape Descriptors
Past Cohort
Clustering
Find clusters with similar patterns
New Patient
Generative Model
Prediction
Series of scans about pelvic region
Rely on cluster of similar patients for predictions

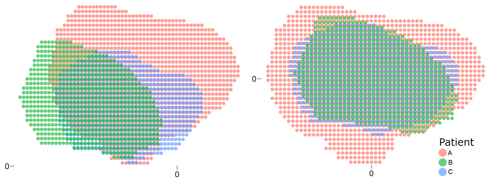
**Figure 1:** *Schematic depiction of the workflow of PREVIS [FMCM*21]; in cyan, we highlight the focus of the current paper.*

to the new incoming patient, based on incomplete data of the latter. For this, several clustering alternatives and different metrics are available—also, influencing the remainder of the workflow. Practically, we are interested in understanding the following two aspects: *What are the effects of using a different clustering methods (e.g., fuzzy or robust methods)?* and *What are the effects of using different parametrizations in the clustering (e.g., different similarity measures)?* To this end, we will investigate the influence of different settings on the clusters and predictions, as well as several other algorithms that could provide robust alternative solutions.

## 2. Related Work

To analyze and visualize the variability of organs, 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. Other publications focused specifically on the relationship between organ shape variability and segmentation errors yielded by different algorithms [RBGR18]. 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. To describe the variability of pelvic organs specifically, recent research studies proposed a series of approaches. Raidou et al. [RCMA*18] focused exclusively on the bladder and Furmanová et al. [FGM*20] included a more detailed analysis of the correlation between shape variations and possible toxicities in the entire pelvis. Finally, in *PREVIS* by Furmanová et al. [FMCM*21], a prediction workflow was proposed, where retrospective patients were grouped based on their overall pelvic organ shape variability and their information was used to make predictions for new incoming patients. Although all these 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.

Regarding clustering analysis, several previous works have focused on exploration of various methods specifically for medical data, since clustering plays a crucial role in cohort analysis. Gotz et al. [GSCE11] presented DICON, a tree-map based interactive visualization tool to analyze patient clusters based on their electronic health records. Klemm et al. [KLR*13] evaluated different proximity measures for clustering of spinal canals. Glaßer et al. [GNPS13] applied various clustering methods to medical data to classify tumors. Meuschke et al. [MLK*16] investigated the performance of different clustering algorithms on aortic blood flow. These works present diverging results, demonstrating that a thorough evaluation is necessary for each application scenario.



(a) Prostate-centered (b) Individually-centered

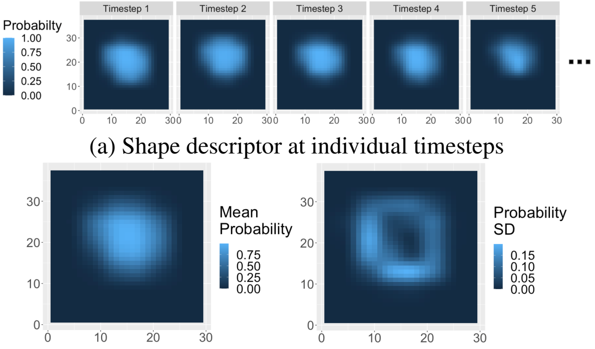
**Figure 2:** *Bladder centering comparison for three patients.*

## 3. Method

**Expression of organ shapes by descriptors:** In order to describe the variability of the organs in a mathematical way, a quantitative representation is required. Shape description methods are applied to generate one-dimensional vectors, capturing the shape of the organs. We are looking for a shape description method that should be fully reversible to the initial shape, extensible to new patients, controllable in terms of size, and enabling easy comparison of organ shapes. We start the construction of the descriptors by calculating the *bounding box* that covers the organ space in an entire cohort. The bounding box is then sampled with *target points* distributed in a regular grid which forms the basis of the descriptor. 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. Next, a value representing the *probability of presence* of the organ is assigned to each of the target points. *PREVIS* relied on a kernel density estimation-based approach proposed by Akgül et al. [ASYS06]. Alternatively, we divide the organ space into grid of cells, where each target point forms the center of one cell. Next, the probability at the target point is computed as the percentage of overlap between the cell and the original organ. We call this a *resolution-based descriptor* as it essentially reduces the resolution of the input shape. Finally, the target points are ordered into a one-dimensional vector, which is practical for further calculations. When we need to reconstruct the original shape, we employ *upsampling* of the descriptor vector. First, we reconstruct the 3D grid of probability values from the vector. Then, we add new rows and columns between the existing ones and use linear interpolation to impute the new values. This process can be repeated iteratively to reach a desired resolution. The upsampled values are smoothed, and finally, a cut-off value is set, above which a target point is considered part of the output shape.

To ensure that CT scans from different patients and time steps are comparable, each CT scan is *centered* with respect to the prostate. Thus, the bladder and rectum are analyzed in terms of their relative position to the prostate. However, this relative location might differ across patients. For the analysis and prediction of individual organs, improvements might be gained by centering each organ type separately. In this way, each organ would 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 2.

Figure 3 (a) highlights the changes throughout a number of timesteps during treatment for a sample patient's bladder. To de-



(a) Shape descriptor at individual timesteps

(b) Mean descriptor

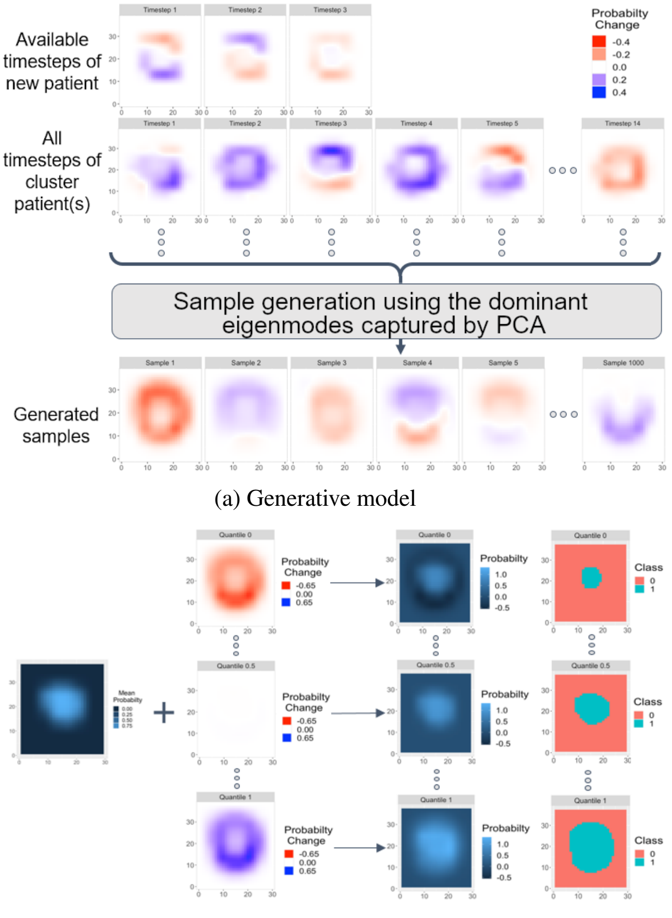
(c) Standard deviation

**Figure 3:** *Bladder shape descriptor computation for one patient.*

scribe the overall patterns of each patient, an *aggregation* of their timestep descriptors is performed, using the mean and standard deviation of the probability at each target point for each individual organ (Figure 3 (b,c)). Finally, the mean and standard deviation descriptors of all organs are concatenated into one single *patient descriptor*. Since all patients have the same descriptor length, ensured by the use of same target point base, they can be easily compared and used as an input for the patient clustering.

**Clustering of past patients:** In our workflow, the predictions for new patients are constrained by the information coming from a group of past patients with the highest similarity. To facilitate this, *clustering techniques* are used to divide the patient cohort according to their similarity. One option for a clustering algorithm is an agglomerative hierarchical clustering (AHC), which was also used in *PREVIS*. Other clustering alternatives that we include in our approach are $k$-means, $k$-medoids, model-based clustering, and fuzzy $c$-means, being the most common methods. In addition to the algorithm choice, we also need to decide on a *distance measure* and a *linkage criterion* for the AHC. The distance measure specifies a mathematical measure to describe the distance between two observations and the linkage method determines which observations are used to compare different clusters. *PREVIS* relied on Euclidean distance and complete linkage, and we additionally include Manhattan, Minkowski, Canberra, Binary and Maximum distance, as well as single, average, McQuitty, median, centroid, and Ward linkage.

**Prediction workflow:** The information of the new patient and of the most similar cluster of patients is then used to generate a large sample set of possible shape variations. For this, the *probability change* across different timesteps is calculated as a deviation from the mean shape of each patient. Subsequently, the main patterns from the shape variations are extracted by calculating the dominant eigenmodes of the data using principal component analysis (Figure 4 (a)). This information is applied to generate further samples of the shape changes. A *user-selected quantile* of the changes at each specific target point is then calculated. The resulting values represent a deformation that is added to the mean shape of the new patient, thereby forming the final prediction. Quantiles below 0.5 consist of negative changes in the probability and result in a shrinkage of the mean shape. Quantiles above 0.5 represent a shape variation that increases the mean shape in size. An example of a bladder shape prediction including the probability change for different quantiles and a cut-off of 0.5 is shown in Figure 4 (b).



(a) Generative model

(b) Shape prediction

**Figure 4:** *Prediction workflow: Shape deformation samples generated from the past patient data using PCA (a) are applied to the new patient (b, left) to predict shape variations (b, right).*

## 4. Results

**Expression of organ shapes by descriptors:** To evaluate the ability of different descriptors to reconstruct the original input shape, we first focus on the probabilistic shape description method used in *PREVIS*. Figure 5 (a) presents a comparison of the upsampled shape descriptors using various cut-off values and their respective reconstruction accuracy measured by the Dice coefficient. Each line in this visualization represents a single patient, with the results calculated as an average accuracy of the available timesteps for a given organ. In this case, different patients require significantly different cut-off settings to obtain the best reconstruction of their shape descriptor. Also, 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 (a). However, this was expected, due to the more irregular and non-spherical shape of the rectum, compared to the other two organs.

Performing the same analysis for our resolution-based descriptors, we get much better results, as shown in Figure 5 (b). Apart from more consistent patterns, this method also achieves an overall better reconstruction accuracy. The peak accuracy across all three organ types is achieved by a cut-off value of 50%. 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. The rectum tends to yield the worst reconstruction accuracy, due to its irregular shape. Therefore, opting for smaller grid dimensions and increasing the overall number of target points might improve the quality of the descriptors and their reconstructions. Figure 5 (c) shows this by decreasing the dimensions of the grid from 15 to 10 *mm*. Compared to Figure 5 (b), the reconstruction accuracy has increased for all organs. In addition, the rectum now shows more consistent patterns across patients.

■ Overall, our analysis has highlighted that ***our resolution-based descriptors provide an alternative solution with more accurate and reliable organ representations*** than *PREVIS*—especially with a fine grid resolution (e.g., 10 *mm* grid resolution).

**Clustering of past patients:** For the topic of clustering, we first focus on hierarchical clustering, which was also the method of choice in the original implementation of *PREVIS*. This approach requires two further decisions. 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 are compared and subsequently merged into larger clusters.

Focusing on the calculation of *distance*, 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 of the shape descriptors for each individual patient. 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. 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. Figure 6 illustrates the mean overlap for the bladder using different distance measures. It also compares the prostate- and individually-centered descriptors of Figure 2. Overall, the individually-centered descriptors perform better and the desired ranking patterns are best achieved by the Euclidean distance.

The *linkage method* determines how the observations captured in the distance metric are organized into a hierarchy of clusters. To explore this topic in isolation, we restrict the input distance metric to the individually-centered descriptors and the euclidean distance as the distance measure, as this combination has been shown to be optimal. The key question for this problem is how many clusters should be chosen. To address this, we compared the elbow plots for different linkage methods. Our results showed, that depending on the linkage method used, the use of 3 or 4 clusters may be a good choice. The only exception to these patterns is the average linkage, which tends to produce irregular results.

To examine the underlying cluster assignments under different linkage methods, Figure 7 visualizes the patient cohort using multidimensional scaling and compares the cluster assignment of individual patients using 3 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. For all other linkage methods, the clusters are divided into more equal groups of patients, especially for the complete, McQuitty, and Ward methods. AHC provides a way to cluster patients, but *other clustering methods* can be used. Among all alternatives we investigated ($k$-means, $k$-medoids, model-based, fuzzy $c$-means, and AHC), the clustering results did not differ significantly.

■ Overall, our analysis has highlighted that ***different settings (distance metric and linkage) yield only slightly different clusters.***

**Prediction worfklow:** To evaluate the outcomes of the prediction workflow under different settings, we employ a leave-one-out approach by simulating each patient as a new patient with incomplete data—similarly to the evaluation presented in *PREVIS*. One general finding is that when more variation is considered for the predictions, larger deviations can be observed across different patients and clustering settings. However, even for the most extreme cases of added variation, there is little evidence that different clustering settings influence the outcomes of the prediction. Figure 8 shows the prediction results achieved with AHC with Euclidean distance for the different linkage methods shown in Figure 7. We can see that the linkage methods of centroid, median, and single perform slightly worse than the other methods. However, the difference is not large enough to draw general conclusions about the superiority of the other settings.

■ Overall, our analysis has highlighted that different clustering methods provide comparable results, although ***when more shape variability is considered the outcomes start to deviate.***

## 5. Conclusions and Future Work

In this work, we analyze the impact of statistical and machine learning choices on generative models that predict pelvic organ variability for RT. We particularly focused on the organ shape descriptors and clustering methods (and their parametrizations). Our results indicate that the most significant choice is the shape descriptor. In the future, we would like to investigate the influence of missingness and/or noise on the input data, as well as the propagation of uncertainty throughout the predictive workflow.

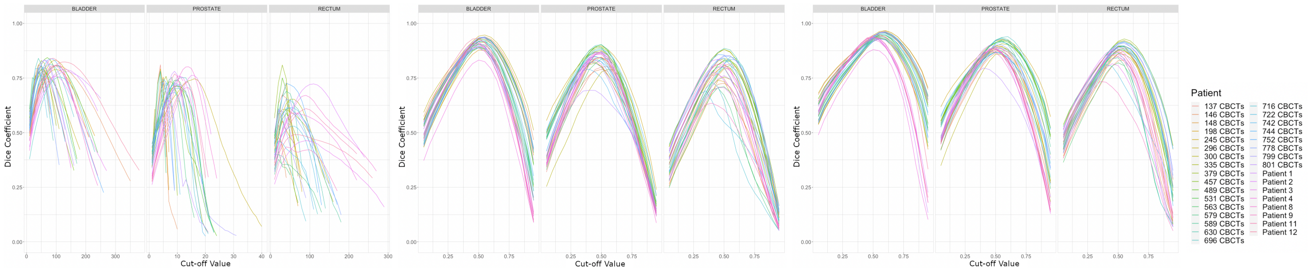
## References

[ASYS06] Akgül C. B., Sankur B., Yemez Y., Schmitt F.: A framework for histogram-induced 3D descriptors. In *2006 14th European Signal Processing Conference* (2006), IEEE, pp. 1–5. 2

[BBP10] Busking S., Botha C. P., Post F. H.: Dynamic Multi-View Exploration of Shape Spaces. *Computer Graphics Forum 29*, 3 (2010), 973–982. 2

[FGM*20] Furmanová K., Grossmann N., Muren L. P., Casares-Magaz O., Moiseenko V., Einck J. P., Gröller M. E., Raidou R. G.: VAPOR: Visual Analytics for the Exploration of Pelvic Organ Variability in Radiotherapy. *Computers & Graphics 91* (2020), 25–38. 1, 2

[FMCM*21] Furmanová K., Muren L. P., Casares-Magaz O., Moiseenko V., Einck J. P., Pilskog S., Raidou R. G.: PREVIS: predictive visual analytics of anatomical variability for radiotherapy decision support. *Computers & Graphics 97* (2021), 126–138. 1, 2



(a) *PREVIS* descriptors (15 mm grid resolution)

(b) Resolution-based descriptors (15 mm grid resolution)

(c) Resolution-based descriptors (10 mm grid resolution)

**Figure 5:** *Reconstruction accuracy measured by the Dice coefficient for different shape descriptors and/or grid resolutions. Resolution-based descriptors with a finer grid (c) provide more accurate organ representations.*

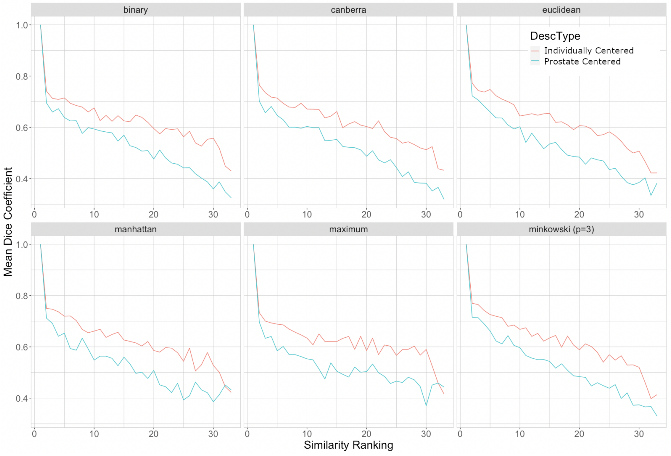


**Figure 6:** *Mean overlap of bladders w.r.t. distance metric ranking. The Euclidean distance exhibits the most stable linear relationship between organ similarity and overlap.*

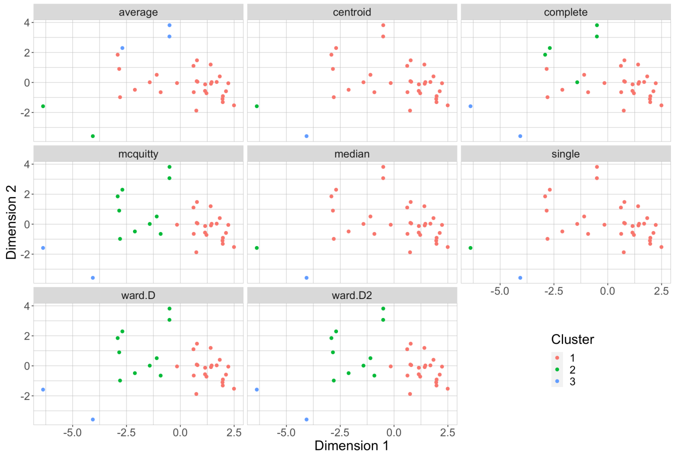


**Figure 7:** *Patient assignment using 3 clusters for different linkages. Most methods tend to merge patients into one single cluster.*

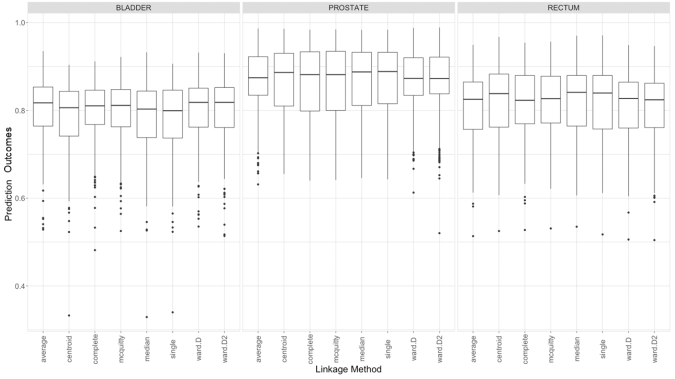


**Figure 8:** *Predictions achieved by different linkage methods for a variation quantile of 1 (using Euclidean distance). No significant impact of the linkage method on the prediction can be observed.*

[GNPS13] Glasser S., Niemann U., Preim B., Spiliopoulou M.: Can we distinguish between benign and malignant breast tumors in DCE-MRI by studying a tumor's most suspect region only? In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (2013), IEEE, pp. 77–82. 2

[GSCE11] Gotz D., Sun J., Cao N., Ebadollahi S.: Visual cluster analysis in support of clinical decision intelligence. In *AMIA Annual Symposium Proceedings* (2011), American Medical Informatics Association, p. 481. 2

[KLR*13] Klemm P., Lawonn K., Rak M., Preim B., Tönnies K. D., Hegenscheid K., Völzke H., Oeltze S.: Visualization and analysis of lumbar spine canal variability in cohort study data. In *Vision, Modeling & Visualization* (2013), The Eurographics Association, pp. 121–128. 2

[MLK*16] Meuschke M., Lawonn K., Köhler B., Preim U., Preim B.: Clustering of aortic vortex flow in cardiac 4D PC-MRI data. In *Bildverarbeitung für die Medizin 2016* (2016), Springer Berlin Heidelberg, pp. 182–187. 2

[RBGR18] Reiter O., Breeuwer M., Gröller M. E., Raidou R. G.: Comparative Visual Analysis of Pelvic Organ Segmentations. In *EuroVis 2018 - Short Papers* (2018), The Eurographics Association, pp. 37–41. 2

[RCMA*18] Raidou R. G., Casares-Magaz O., Amirkhanov A., Moiseenko V., Muren L. P., Einck J. P., Vilanova A., Gröller M. E.: Bladder Runner: Visual Analytics for the Exploration of RT-Induced Bladder Toxicity in a Cohort Study. *Computer Graphics Forum 37*, 3 (2018), 205–216. 1, 2

[SRM*19] Schlachter M., Raidou R. G., Muren L. P., Preim B., Putora P. M., Bühler K.: State-of-the-Art Report: Visual Computing in Radiation Therapy Planning. In *Computer Graphics Forum* (2019), vol. 38, pp. 753–779. 1