Detecting clinically meaningful shape clusters in medical image data: metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches

Jan L. Bruse, Maria A. Zuluaga, Abbas Khushnood, Kristin Mcleod, Hopewell N. Ntsinjana, Tain-Yen Hsia, Maxime Sermesant, Xavier Pennec, Andrew M. Taylor, Silvia Schievano

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

Detecting Clinically Meaningful Shape Clusters in Medical Image Data: Metrics Analysis for Hierarchical Clustering applied to Healthy and Pathological Aortic Arches

Jan L Bruse*, Maria A Zuluaga, Abbas Khushnood, Kristin McLeod, Hopewell N Ntsinjana, Tain-Yen Hsia, Maxime Sermesant, Xavier Pennec, Andrew M Taylor, and Silvia Schievano; for the Modelling of Congenital Hearts Alliance (MOCHA) Collaborative Group

Abstract*—*Objective: **Today's growing medical image databases call for novel processing tools to structure the bulk of data and extract clinically relevant information. Unsupervised hierarchical clustering may reveal clusters within anatomical shape data of patient populations as required for modern Precision Medicine strategies. Few studies have applied hierarchical clustering techniques to three-dimensional patient shape data and results depend heavily on the chosen clustering distance metrics and linkage functions. In this study, we sought to assess clustering classification performance of various distance/linkage combinations and of different types of input data to obtain clinically meaningful shape clusters.** ***Methods:*** **We present a processing pipeline combining automatic segmentation, statistical shape modelling and agglomerative hierarchical clustering to automatically subdivide a set of 60 aortic arch anatomical models into healthy controls, two groups affected by congenital heart disease, and their respective subgroups as defined by clinical diagnosis. Results were compared with traditional morphometrics and principal component analysis of shape features.** ***Results:*** **Our pipeline achieved automatic division of input shape data according to primary clinical diagnosis with high F-score (0.902±0.042) and Matthews Correlation Coefficient (0.851±0.064) using the Correlation/Weighted distance/linkage combination. Meaningful subgroups within the three patient groups were obtained and benchmark scores for automatic segmentation and classification performance are reported.** ***Conclusion:*** **Clustering results vary depending on the distance/linkage combination used to divide the data. Yet, clinically relevant shape clusters and subgroups could be found with high specificity and low misclassification rates.** ***Significance:*** **Detecting disease-specific clusters within medical image data could improve image-based risk assessment, treatment planning and medical device development in complex disease.**

Asterisk indicates corresponding author

J. L. Bruse, A. Khushnood, H. N. Ntsinjana, T.-Y. Hsia, A. M. Taylor and S. Schievano are with the Centre for Cardiovascular Imaging, University College London, Institute of Cardiovascular Science & Cardiorespiratory Unit, Great Ormond Street Hospital for Children, London, UK (email: jan.bruse.12@ucl.ac.uk)

M. A. Zuluaga is with the Translational Imaging Group, Centre for Medical Image Computing, University College London, London, UK

K. McLeod is with the Simula Research Laboratory, Cardiac Modelling Department, Oslo, Norway, and KardioMe s.r.o, Bratislava, Slovakia

M. Sermesant and X. Pennec are with Université Côte d'Azur, Inria Sophia Antipolis-Méditeranée, ASCLEPIOS Project, Sophia Antipolis, France

MOCHA Collaborative Group: see Acknowledgment.

***Index Terms*—Aortic arch, automatic segmentation, cardiovascular magnetic resonance imaging, clinical decision support, congenital heart disease, hierarchical clustering, statistical shape analysis**

I. Introduction

Modern medical imaging techniques such as computed tomography (CT) and magnetic resonance (MR) imaging provide detailed and accurate anatomical and functional information of inner body structures and organs, making them widely used tools for diagnosis and treatment planning. Consequently, medical image databases are growing and valuable patient data are accumulating, calling for novel approaches to process and extract clinically relevant information not only on a case-by-case basis, but also considering entire patient populations [1], [2], [3].

Many computational image processing pathways focus on segmentation of body structures [4], [5] or apply classification algorithms to automatically distinguish between healthy and disease [6], [7], [8]. Yet, to date few studies have looked at tools that can be applied *after* those two crucial steps, computational tools that can help understand a disease once anatomical shape information is given and once a diagnosis has been made. Automated clustering techniques from the field of data mining have been widely used in genomics, taxonomy and chemoinformatics to structure large amounts of data into subgroups, thereby revealing previously unknown, yet relevant patterns within a given population [9], [10]. We believe that such an approach may prove beneficial as well for the analysis of complex three-dimensional (3D) anatomical models from medical image data in order to close the gap between mere data and useful knowledge, as desired in current Precision Medicine or "Precision Imaging" approaches [3]. Clinical image assessment of inner body structures usually reveals a patient's dominant pathology, but it often remains unclear how individual image data relate to other patients with the same disease or primary diagnosis. Grouping patients according to anatomical similarity and taking into account clinical history and other functional or outcome parameters may ultimately allow refined, cluster-adapted treatment and follow-up strategies and could assist in risk-stratification when scanning a new patient with similar diagnosis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

Hierarchical clustering techniques seem to be an attractive way to discover anatomical subgroups from medical image data as they are inherently unsupervised, thus do not require any prior information about the study population and, unlike K-means clustering, do not require specifying an expected number of subgroups [11], [12], [13]. Furthermore, clustering results can be graphically summarised in a *dendrogram* that depicts in a tree-like diagram how similar subjects are grouped together, while dissimilar subjects are placed on different branches of the tree. However, evaluation of subject similarity or dissimilarity and clustering results heavily depends on the choice of both similarity or distance metric (with low inter-subject distance relating to higher similarity) and linkage function determining how subjects are linked together to form a subgroup [12], [13]. Depending on the chosen distance/linkage combination, clustering results may vary substantially – potentially rendering meaningless results [14], [15]. While previous studies have analysed clustering techniques based on generic shapes or two-dimensional (2D) shape data [16], few have assessed hierarchical clustering performance using actual patient data in a realistic setting, i.e. using three-dimensional (3D) anatomical models of healthy and pathological shapes derived from medical images [17], [18]. In general, medical image hierarchical clustering performance data including validation against known and clinically relevant clusters are sparse. In this study, we aimed to investigate whether and how hierarchical clustering can be used to automatically divide a bulk of unlabelled clinically acquired cardiovascular magnetic resonance (CMR) image data into clusters and subgroups that could be of clinical relevance.

Specifically, we sought to analyse clustering classification performance of various distance/linkage combinations applied to a population of 60 aortic arch anatomical models, automatically segmented from CMR data, composed of three equally-sized subgroups of healthy aortic arches, arches post aortic coarctation repair (COA) [19] and arches post arterial switch operation (ASO) [20]. COA and ASO patients suffer from congenital heart disease (CHD), which manifests itself in abnormalities of cardiovascular structures (here, the aorta, known to present shape patterns abnormal from healthy individuals [19], [21] [22]). Anatomy plays a crucial role in both diagnosis and therapy of CHD, as shape abnormalities often lead to functional impairment, requiring intervention. COA and ASO image data provide an excellent platform to test unsupervised clustering algorithms, as newly found shape clusters or subgroups within those diseases may ultimately impact on novel diagnosis and treatment strategies.

To assure "meaningfulness" (here, clinical relevance) of unsupervised clustering results, we externally validated [15], [14] our results against clinical expert opinion, traditional morphometric parameters and 3D shape analysis via principal component analysis (PCA). We aimed to find the distance metric/linkage function combination that achieved highest classification performance, i.e. that was able to automatically divide the bulk CMR input data into the three clinically meaningful clusters of CTRL, COA and ASO arch shapes with low misclassification rates.

Furthermore, we hypothesised that such clinically meaningful clustering on a macrolevel yields meaningful shape subgroups (i.e. "clusters within clusters") on lower-level hierarchies of the clustering tree as well, which may allow the detection of novel disease patterns in future studies.

II. Methods

The study outline is as follows: all aortic arch shape models were automatically segmented from CMR data and were parameterised within one common mathematical framework using a non-parametric statistical shape modelling (SSM) approach based on non-rigid registration of a computed template shape [23], [24]. Based on this shape data, we applied principal component analysis (PCA) for more detailed assessment of 3D shape features prior to cluster analysis. Hierarchical clustering was then performed on both the full, unprocessed shape data and the reduced PCA dataset to determine the input and the distance/linkage combination yielding clustering closest to the clinical expert diagnosis with high classification performance. Lastly, the distance/linkage setting yielding the most meaningful division of the data (with highest *F-score* and *Matthews Correlation Coefficient*) was analysed in more detail.

A. Patient Population

A total of 60 patients, who underwent routine CMR examination (whole heart 3D balanced, steady-state free precession acquisition; 1.5T Avanto MR scanner, Siemens Medical Solutions, Erlangen, Germany) at Great Ormond Street Hospital for Children (GOSH, London, UK) were retrospectively included in the study. The cohort was divided into three subgroups according to their clinical primary diagnosis: 20 healthy subjects whose aortic arch shapes were reported as normal at cardiac assessment (control group CTRL, age 15.2±2.03 years, 3 female), 20 patients who had undergone surgical aortic arch reconstruction for treatment of coarctation of the aorta (COA, 23.1±7.35 years, 4 female) and 20 patients who had their aorta pushed back posteriorly in the Lecompte [20] manoeuvre for arterial switch operation (ASO, 14.4±2.48 years, 4 female). Ethical approval was obtained by the Great Ormond Street Institute of Child Health/GOSH Research Ethics Committee and all patients or legal parent or guardian gave informed consent for research use of the image data.

B. Segmentation and Registration

The aorta including the left ventricle (LV) was segmented automatically using a multi-atlas propagation segmentation approach that applies a locally normalised cross correlation (LNCC) based ranking combined with a consensus based region-of-interest selection, which has been successfully applied to whole heart [25] and right ventricle segmentation from CMR data [4]. For each group, a *leave-one-out* strategy was followed, where 19 manually labelled atlases of the respective group were used to segment one unseen subject, and dice similarity coefficients (DSC) were computed to quantify automatic segmentation accuracy following $DSC=2AB/(A+B)$, where A is the obtained segmentation and B the corresponding ground truth. Automatic segmentation results were visually inspected and, if necessary, manually edited (i.e. cleaned up and improved) using ITKSnap [26].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

Segmentation labels were exported as 3D computational surface meshes in the Visualization Toolkit (VTK) format [27] and visualised in ParaView [28]. All models were cut consistently below the aortic root and at the level of the diaphragm using The Vascular Modeling Toolkit (VMTK, [29]) cutting tools, whilst coronary arteries and head and neck vessels were cut off as close as possible to the arch. All surface meshes were then rigidly registered to one healthy CTRL subject using an Iterative Closest Point (ICP) algorithm [30] prior to template computation (i.e. 3D population anatomical mean shape, see section C). In order to remove bias due to misalignment of input shapes, a *Generalised Procrustes Analysis* (GPA) was adopted, by computing an initial template, realigning the input shapes to the new template via ICP registration and recomputing the template until convergence, as described in [31].

C. Template and Deformation Matrix Computation

The 60 aligned arch surface meshes constituted the input for the template computation using the openly available *Deformetrica* code framework (*www.deformetrica.org*) [32]. The framework computes the 3D template shape of an input shape population, without assuming any point-to-point correspondence between input meshes. This is achieved by modelling shapes as mathematical currents (surrogate representations of shapes), which characterise a shape as a distribution of shape features rather than its actual point coordinates in space [23], [24]. Template and resulting template-dependent shape parameterisations were computed following protocols detailed in [31]. Surface meshes were transferred into a vector space of currents W, generated by a Gaussian kernel. The standard deviation of the kernel λ_W allows control of the currents resolution and was set to 5mm. The template and its transformations φ_i registering template to each subject shape were computed simultaneously using the large deformation diffeomorphic metric mapping (LDDMM) framework [33]. The transformation functions φ_i were defined within another Gaussian kernel vector space V with standard deviation λ_V, set to 20mm, controlling the transformation stiffness. All 3D shape features present in the population were thus encoded by subject-specific transformations of the template φ_i, which are parameterised by a unique set of deformation vectors β_i for each patient shape. Setting λ_W to 5mm and λ_V to 20mm, resulted in a set of 300 β_i per patient. With each β_i having an x, y and z entry, a final *deformation matrix* D_{Full} of size N x n with N=60 included subjects and n=900 deformation momenta comprised all 3D shape information of the input population and was used for further analysis via PCA and hierarchical clustering.

D. Morphometric Analysis and Principal Component Analysis

To investigate whether arch shape characteristics related to *size* and *shape* were sufficiently different between the three groups (i.e. whether the three patient groups translate into three *shape* groups), traditional morphometric analysis was carried out in 2D and in 3D, without controlling for size difference as size itself is a descriptor of pathological paediatric patient arch shape as well. In terms of size, aortic arch model volume V, surface to volume ratio S_{Vol} and arch centreline length CL_{length} were derived automatically using VMTK and Matlab (The MathWorks, Natick, MA). As shape parameters, we considered arch centreline tortuosity CL_{tort} [34], ascending to descending aortic arch diameter ratio $D_{asc,desc}$ and arch width T, manually measured on the image slices as described in [19], [31].

Further, we performed PCA on the covariance matrix of the combined deformation vectors β_i [35] to extract PCA *shape modes*, each describing a certain amount of 3D population shape variability as a deformation of the template shape. Each subject deformation φ_i was projected onto each PCA shape mode to obtain the low-dimensional shape vector $\{f_{i,k}\}, k \in [1, m]$ [35] for each shape mode k and subject i, whose entries parameterise the subject-specific PCA loadings. The $\{f_{i,k}\}$ were compared between the three groups CTRL, COA and ASO, and the $\{f_{i,k}\}$ of the first two PCA shape modes were plotted against each other to visualise potential grouping within the input shape data. The first $m=19$ shape modes, explaining 90% of the total shape variability (determined by the proportion of sorted eigenvalues) were selected [36] and the respective $\{f_{i,k}\}$ combined constituted the reduced PCA shape loading matrix D_{PCA} of size N x m, which described 3D population shape features in terms of the lower-dimensional PCA loadings.

E. Hierarchical Clustering

The shape matrices D_{Full} and D_{PCA} constituted the input for the agglomerative hierarchical clustering algorithm (Matlab). Based on a pre-defined distance (i.e. similarity) metric, clusters are formed by grouping subjects with similar features together, while subjects with distinctly different features are placed in other clusters. This unsupervised approach unveils "naturally occurring" subgroups within the data, without depending on prior user input [12], [16]. Here, features of interest were 3D aortic arch shape features, parameterised by the entries of D_{Full} and D_{PCA}. The algorithm can be described as follows [37], [38]:

1) Compute distances between every pair of subjects within the input dataset to obtain a metric of pairwise subject similarity (treating each subject as its own cluster).
2) Form binary cluster from two closest (most similar) subjects (using distance metric) or clusters (using linkage function).
3) Recompute distances between newly formed cluster and remaining subjects or clusters.
4) Return to Step 2 until all subjects are included in one large cluster, formed by a tree-like multi-level network of subclusters (dendrogram). At the lowest level, each subject forms its own cluster.
5) Cut off dendrogram branches at a specified level of the hierarchy to assign subjects below each cut to a specific cluster, generating partitions of the data.

To compute pairwise distances between the 60 patient shapes parameterised by deformation row vectors of D_{Full} or PCA shape vectors of D_{PCA}, the following commonly used distance (similarity) metrics *dist* between the vector pair x_s and x_t were computed (with D being of size N x n, with N (1-by-n)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

row vectors $x_i, i \in [1, N]$; for D_{Full} with $n \in [1, \dots, 900]$ and for D_{PCA} with $n \in [1, \dots, 19]$) [38]:

$$dist_{Euclidean} = \sqrt{\sum_{j=1}^{n} \left| x_{sj} - x_{tj} \right|^2} \triangleq \left\| x_{sj} - x_{tj} \right\|_2 \quad (1)$$

$$dist_{\text{StandardisedEuclidean}} = \sqrt{\sum_{j=1}^{n} \frac{\left| x_{sj} - x_{tj} \right|^2}{s_j^2}} \quad (2)$$

with s_j being the standard deviation of the x_s and x_t over the sample set.

$$dist_{Cityblock} = \sum_{j=1}^{n} \left| x_{sj} - x_{tj} \right| \triangleq \left\| x_{sj} - x_{tj} \right\|_1 \quad (3)$$

$$dist_{Chebychev} = \max_j \left\{ \left| x_{sj} - x_{tj} \right| \right\} \triangleq \left\| x_{sj} - x_{tj} \right\|_\infty \quad (4)$$

$$dist_{Cosine} = 1 - \frac{x_s x_t'}{\sqrt{(x_s x_s')(x_t x_t')}} \triangleq 1 - \frac{x_s \cdot x_t}{\left\| x_s \right\| \left\| x_t \right\|} \quad (5)$$

$$dist_{Correlation} = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'} \sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'}}$$

$$\text{with } \bar{x}_s = \frac{1}{n} \sum_{j=1}^{n} x_{sj} \text{ and } \bar{x}_t = \frac{1}{n} \sum_{j=1}^{n} x_{tj} \quad (6)$$

$$dist_{Spearman} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'} \sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}}$$

where $\bar{r}_s = \frac{1}{n} \sum_j r_{sj} = \frac{(n+1)}{2}$ and $\bar{r}_t = \frac{1}{n} \sum_j r_{tj} = \frac{(n+1)}{2}$; r_s and r_t are the coordinate-wise rank vectors of x_s and x_t. (7)

After defining a distance metric between *pairs of subject shapes*, a linkage function then uses the generated distance data to join *groups of subjects* together into binary clusters and link those to higher level larger clusters, until all subjects are linked together. The linkage function thus defines the similarity or distance between two groups of subjects and is used to generate the dendrogram. The order in which subjects are clustered together is determined by the type of linkage method. For each distance metric, the following commonly used linkage methods were applied to generate a dendrogram. For subjects or clusters s and t joined into cluster $s \cup t$, the new distance between this cluster and another subject or cluster k is generally defined by the Lance-Williams dissimilarity update formula $link(s \cup t, k)$ (8), which defines different types of linkage methods, depending on the choice of the parameters α_s, α_t, β and γ as follows [10]:

$$link(s \cup t, k) = \alpha_s dist(s,k) + \alpha_t dist(t,k) + \beta dist(s,t) + \gamma \left| dist(s,k) - dist(t,k) \right| \quad (8)$$

$$link(s \cup t, k)_{Average} : \alpha_s = \frac{n_s}{n_s + n_t}, \alpha_t = \frac{n_t}{n_t + n_s}, \quad (9)$$
$$\beta = 0, \ \gamma = 0$$

$$link(s \cup t, k)_{Centroid} : \alpha_s = \frac{n_s}{n_s + n_t}, \alpha_t = \frac{n_t}{n_t + n_s}, \quad (10)$$
$$\beta = -\frac{n_s n_t}{(n_s + n_t)^2}, \ \gamma = 0$$

$$link(s \cup t, k)_{Complete} : \alpha_s = \frac{1}{2}, \alpha_t = \frac{1}{2}, \ \beta = 0, \ \gamma = \frac{1}{2} \quad (11)$$

$$link(s \cup t, k)_{Median} : \alpha_s = \frac{1}{2}, \alpha_t = \frac{1}{2}, \ \beta = -\frac{1}{4}, \ \gamma = 0 \quad (12)$$

$$link(s \cup t, k)_{\text{Single}} : \alpha_s = \frac{1}{2}, \alpha_t = \frac{1}{2}, \ \beta = 0, \ \gamma = -\frac{1}{2} \quad (13)$$

$$link(s \cup t, k)_{Ward} : \alpha_s = \frac{n_s + n_k}{n_s + n_t + n_k}, \alpha_t = \frac{n_t + n_k}{n_s + n_t + n_k}, \quad (14)$$
$$\beta = -\frac{n_k}{n_s + n_t + n_k}, \ \gamma = 0$$

$$link(s \cup t, k)_{Weighted} : \alpha_s = \frac{1}{2}, \alpha_t = \frac{1}{2}, \ \beta = 0, \ \gamma = 0 \quad (15)$$

Note that *dist* can be any of the distance metrics defined in (1-7); n_s, n_k, n_t is the number of subjects in cluster *s*, *k*, *t*, respectively. *Centroid*, *Median* and *Ward* linkage methods are appropriate for *Euclidean* distances only [38]. Cutting the dendrogram horizontally at a particular height or level partitions the data into shape subgroups [12]. As we first aimed to assess whether the clustering algorithm was able to distinguish between CTRL, COA and ASO groups, dendrograms were cut automatically at a level that yielded three large shape clusters.

F. Clustering Classification Performance Measures

Based on the majority of group members associated with one cluster, each cluster was automatically labelled either CTRL ($Class_1$), COA ($Class_2$) or ASO ($Class_3$) and numbers of assigned subjects from each of the three classes were recorded in a *confusion matrix* to assess clustering classification performance. All correctly assigned subjects for each class are shown on the diagonal of the matrix. For each of the three classes $Class_j$ $j \in [1,3]$, the total number of *true positives* (TP_j, e.g. in case of the CTRL class, the actual CTRLs that were correctly classified as CTRL), *false positives* (FP_j, e.g. COA and/or ASO that were incorrectly classified as CTRL), *false negatives* (FN_j, e.g. CTRLs that were incorrectly classified as COA and/or ASO) and *true negatives* (TN_j, e.g. all remaining subjects, correctly classified as non-CTRL) were derived from the confusion matrices.

With these values, overall classification performance was computed using macroaveraging (denoted with subscript *M*) [39] over *L*=3 classes of the following performance measures:

$$Recall_M = \frac{\sum_{j=1}^{L} \frac{TP_j}{TP_j + FN_j}}{L} \quad (16)$$

$$Specificty_M = \frac{\sum_{j=1}^{L} \frac{TN_j}{FP_j + TN_j}}{L} \quad (17)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

$$Precision_M = \frac{\sum_{j=1}^{L} \frac{TP_j}{TP_j + FP_j}}{L} \quad (18)$$

$$Accuracy_M = \frac{\sum_{j=1}^{L} \frac{TP_j + TN_j}{TP_j + FN_j + FP_j + TN_j}}{L} \quad (19)$$

To minimise chance findings and bias associated with those traditional measures, we also computed (macroaveraged) *Informedness*, which relates to the probability that there has been an informed classification as opposed to mere guessing, and *Markedness*, defined as [40]:

$$Informedness_M = \frac{\sum_{j=1}^{L} \frac{TP_j}{TP_j + FN_j} + \frac{TN_j}{FP_j + TN_j} - 1}{L} \quad (20)$$

$$Markedness_M = \frac{\sum_{j=1}^{L} \frac{TP_j}{TP_j + FP_j} + \frac{TN_j}{TN_j + FN_j} - 1}{L} \quad (21)$$

To provide a summary of the above measures, the macroaveraged *F-score*$_M$ (weighted harmonic mean of *Recall* and *Precision)* and *Matthew's Correlation Coefficient (MCC*$_M$*)* (geometric mean of *Informedness* and *Markedness* [41]*)* were computed as follows:

$$F-score_M = \frac{2 Precision_M \, Recall_M}{Precision_M + Recall_M} \quad (22)$$

$$MCC_M = \frac{\sum_{j=1}^{L} \frac{TP_j \, TN_j - FP_j \, FN_j}{\sqrt{(TP_j + FP_j)(TP_j + FN_j)(TN_j + FP_j)(TN_j + FN_j)}}}{L} \quad (23)$$

F-score$_M$ and *MCC*$_M$ scores were used to evaluate overall classification performance of the various distance metric and linkage combinations. Note that *F-score* values range from 0 for worst to 1 for best classification performance, whereas MCC ranges from -1 for total disagreement over 0 for random guessing to +1 for perfect prediction of classes [41]. In the following, the qualitative term "best" refers to highest possible classification performance in terms of both *F-score*$_M$ and *MCC*$_M$ score being close to the value 1.

G. Validation of Clustering Results

Clustering results were evaluated using *10-fold cross validation* (CV), leaving out *N*/10 randomly selected subjects, and recomputing template, D_{Full} and D_{PCA}, until each subject had been left out once. Classification performance measures were calculated for each of the 10 CV runs, looping through all 49 distance metric/linkage combinations for the two different input matrices D_{Full} and D_{PCA}, respectively. All clustering runs were carried out on a 32GB workstation using one 2.3GHz core. The distance/linkage combination with the best classification performance based on mean *F-score*$_M$ and *MCC*$_M$ was chosen for further analyses of the full data matrix, comprising all *N*=60 subjects. Results of this final clustering were visualised as a dendrogram and compared to PCA results.

H. Statistical Analysis

For all analysed size, shape and PCA shape vector entries, mean and 95% confidence intervals (95CIs) based on the patient cohorts are reported. For classification performance measures, mean and 95CIs are reported based on the CV runs.

To compare distributional differences between the three patient groups CTRL, COA and ASO, independent analysis of variance (ANOVA) was performed. Prior to ANOVA, homogeneity of variance was assessed using Levene's test. In case homogeneity of variance was violated, Welch's test was performed. When ANOVA showed significance, post hoc tests were carried out for pairwise group comparisons and Bonferroni adjusted to control for Type I error rates. Statistical significance was assumed at level $p < .05$. All statistical tests were carried out using R v3.3.1 (R Foundation for Statistical Computing, Vienna, Austria).

III. Results

A. Segmentation

Average segmentation runtime was approximately 2 hours per patient (parallel processing on a 24 core, 2.3GHz, 32GB RAM workstation). Average Dice scores (±95CI) for the automatically computed segmentation labels compared to their respective ground truths were 0.917±0.026 for the CTRL, 0.944±0.012 for the COA and 0.913±0.033 for the ASO group. Final automatic segmentation labels required a maximum of 10 minutes manual clean-up.

B. Comparison of Traditional Shape Parameters

In terms of size, significant distributional differences in *V* (Fig. 1a) were found between the COA and CTRL group (p=2e-07), and the COA and ASO group (p=7e-06). S_{Vol} distributions (Fig. 1b) differed significantly between the COA and CTRL group (p=1e-06), and the COA and ASO group (p=3e-03). Distributional differences in CL_{length} (Fig. 1c) were found between the COA and CTRL group (p=5e-05) and the COA and ASO group (p=1e-06). Overall, COA aortic arches were significantly larger and more compact, whereas arch models from the CTRL and ASO group were of similar size. Following this analysis, we would expect the clustering algorithm to confuse CTRL and ASO shapes, while separating out well the COA group, if it mainly took into account size differences between input shapes.

With regard to measured shape parameters, significant differences between all three groups were found for CL_{tort} following post hoc analyses (p=2e-07 for COA vs CTRL, p=1e-14 for COA vs ASO and p=1e-02 for CTRL vs ASO, Fig. 1d). Similarly, $D_{asc,desc}$ distributions (p=1e-05 for COA vs CTRL, p=7e-10 for COA vs ASO and p=4e-02 for CTRL vs ASO, Fig. 1e) and *T* distributions differed significantly (p=6e-08 for COA vs CTRL, p=2e-16 for COA vs ASO and p=3e-09 for CTRL vs ASO, Fig. 1f) between all three groups, with COA arches showing generally more tortuous and wider arch shapes with higher ascending to descending aortic diameter ratios than the other two groups and ASO arches being the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

least wide, least tortuous with the lowest ascending to descending arch diameter ratios.

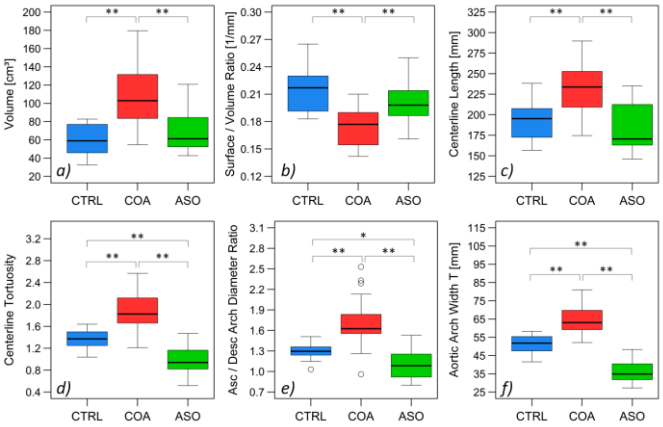

*Fig. 1: Boxplots of size (a-b) and shape (d-f) morphometric parameters describing differences in aortic arch shape between the three patient groups CTRL, COA and ASO. The thick line within the box represents the median value, box height represents the interquartile range and whiskers extend to the maximum and minimum value, respectively. * denotes statistical significance at level p<.05; ** at level p<.01.*

C. Principal Component Analysis of 3D Shape Features

The first three shape modes are visualised in Fig. 2. PCA shape mode 1 accounted for 35.4% of shape variability. It described shape change from an overall small and short, ASO-like arch shape with narrow arch width towards a large, COA-like arch shape with high arch width, dilated root and ascending aorta, and more tortuous descending aorta continuation (Fig. 2a). In terms of $\{f_{i,1}\}$ shape vector entry distributions, COA arches differed significantly from the CTRL group (p=4e-08) and from the ASO group (p=3e-12). CTRL and ASO shape vector entry distributions did not differ significantly (p=.050).

PCA shape mode 2 described shape variability associated with more rounded and wide arches compared to more "gothic" [19] arch shapes with similar arch height but smaller arch width. It accounted for 12.2% of the total shape variability (Fig. 2b). The $\{f_{i,2}\}$ entry distribution for the ASO group differed significantly from the CTRL group (p=2e-08) and from the COA group (p=1e-06), while there was no significant difference between the CTRL and COA groups (p=.930)

PCA shape mode 3 accounted for 7.4% of shape variability. It varied from arch shapes with lower arch height and slightly dilated root to arches with higher arch height but similar arch width and few diameter changes along the arch (Fig. 2c). For this mode, the $\{f_{i,3}\}$ distribution of the CTRL group was significantly different from the COA group (p=2e-02) and from the ASO group (p=3e-03) but no significant difference was found between the COA and ASO group (p=.999)

Following the analysis of traditional shape parameters and the first three PCA shape modes, we concluded that all three selected patient groups were sufficiently different from each other, thus forming three distinct shape groups to be found by the clustering algorithm. Furthermore, plotting the $\{f_{i,1}\}$ and $\{f_{i,2}\}$ for PCA shape modes 1 and 2 against each other revealed a good split between the three groups in PCA 3D shape space (Fig. 2d), justifying the assumption of three large shape clusters within our cohort of 60 patients.

D. Determining best performing Input and Distance/Linkage Combination

Macroaveraged classification performance measures *F-sore*$_M$ and MCC_M for various distance/linkage combinations and the input datasets $D_{Full,CV}$ and $D_{PCA,CV}$ are shown in Fig. 3.

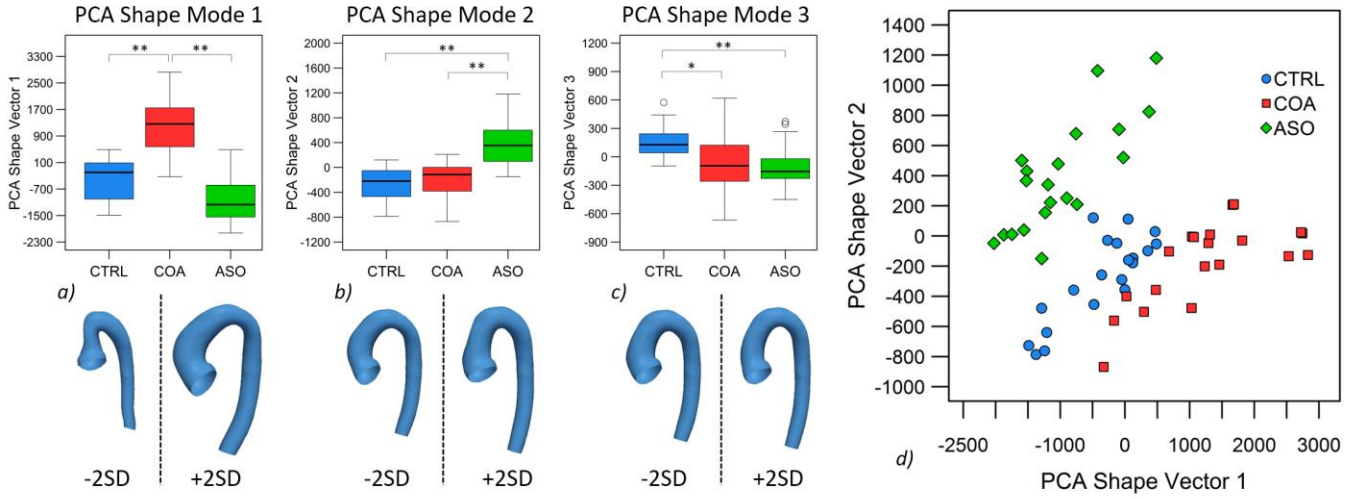

Fig. 2: Results from principal component analysis (PCA) of the deformation shape data D_{Full}. Graphs show boxplots of subject-specific shape vector entries associated with the first three PCA shape modes accounting for 35.4% (mode 1), 12.2% (mode 2) and 7.4% (mode 3) of total shape variability (a-c). For PCA shape mode 1, high shape vector entries were associated with the COA group and with COA-like 3D aortic arch shape features, visualised as a deformation of the computed template shape from -2 to +2 standard deviations (SD) below the graph (a). PCA shape mode 2 related to shape features associated with the ASO group, showing a slightly squeezed, gothic-type arch with dilated aortic root compared to a more rounded arch shape for low shape vector entries (b). PCA shape mode 3 visualised shape changes towards an overall slim aortic arch with relatively constant arch diameter, associated with high shape vector entries and thus the CTRL group (c). The PCA revealed significant differences in 3D arch shape features between the three patient groups. The scatterplot of subject-specific PCA shape vector 1 entries vs PCA shape vector 2 entries (d) revealed grouping among aortic arch input shapes in PCA shape space according to the deformation shape data (denotes statistical significance at level p<.05; ** at level p<.01.).*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

Note that only the linkage option which achieved highest *F-score*$_M$ and *MCC*$_M$ score is shown for each distance metric. Best performing linkages were the same for $D_{Full,CV}$ and $D_{PCA,CV}$, except in the cases of *Cosine* and *Chebychev* distance metrics, where $D_{PCA,CV}$ achieved higher scores using the *Average* linkage instead of *Weighted* linkage function.

In a one-to-one comparison, clustering using $D_{Full,CV}$ yielded better classification performance both in terms of *F-score*$_M$ and *MCC*$_M$ than achieved with $D_{PCA,CV}$. Only the *Chebychev* and *Cityblock* distance metrics performed better for $D_{PCA,CV}$, yet scoring on average below 0.7 for *F-score*$_M$ and below 0.6 for *MCC*$_M$. The worst performance was found for the *Standardised Euclidean* distance, even yielding negative (i.e. highly confused) results in terms of *MCC*$_M$.

On average, the best performing distance metrics (average *F-score*$_M$ above 0.7 and average *MCC*$_M$ above 0.5) were the *Spearman*, *Correlation* and *Cosine* metrics in combination with the *Weighted* linkage and the *Euclidean* distance in combination with the *Ward* linkage. However, particularly *MCC*$_M$ scores revealed weaknesses such as large 95CIs for the *Cosine* metric, making it the most unreliable distance metric. Instead, *Spearman/Weighted*, *Correlation/Weighted* and *Euclidean/Ward* combinations performed consistently well, with the *Correlation/Weighted* combination achieving on average the best classification performance with *F-score*$_M$=0.902±0.042 and *MCC*$_M$=0.851±0.064 for $D_{Full,CV}$. Therefore, the *Correlation/Weighted* distance/linkage combination applied to the full dataset $D_{Full,CV}$ was found to yield the best overall shape clustering results with respect to the three patient groups and was chosen for further analysis.

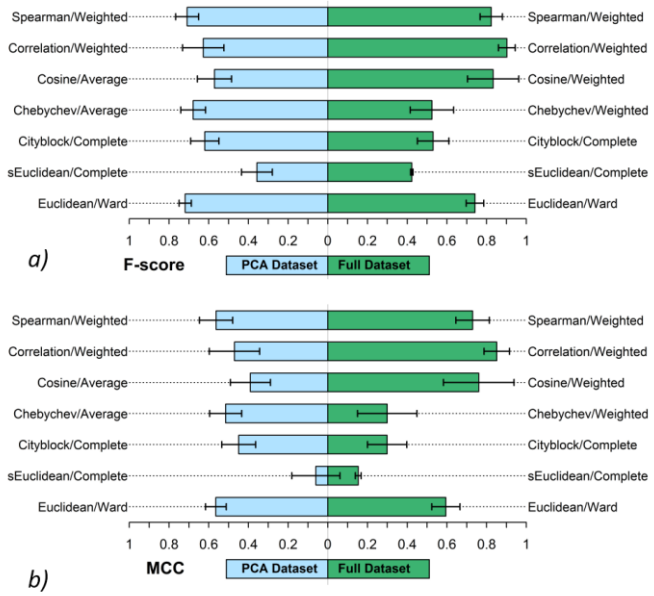

Fig. 3: Clustering classification performance measures for full input dataset ($D_{Full,CV}$, right, green) and reduced PCA shape loading dataset ($D_{PCA,CV}$, left, blue), showing mean and 95CIs of macroaveraged F-score$_M$ *(a) and MCC*$_M$ *(b) for the respective best distance/linkage combinations over 10 CV runs. Overall, the $D_{Full,CV}$ input dataset performed better than the reduced PCA dataset and Spearman/Weighted, Correlation/Weighted and Euclidean/Ward distance/linkage combinations were found to yield good and reliable clustering classification performance.*

E. Analysis of Best Performing Distance/Linkage Combination

Looking at individual classification performance metrics, the *Correlation/Weighted* distance/linkage combination performed consistently well, with average *Informedness*$_M$, *Markedness*$_M$ and *MCC*$_M$ scores above 0.8 and *Specificity*$_M$, *Recall*$_M$, *Precision*$_M$, *F-score*$_M$ and *Accuracy*$_M$ measures around 0.9 (Fig. 4). Highest scores were achieved for *Specificity*$_M$ (i.e. proportion of patients correctly identified as *not* being a member of one of the three groups) with 0.948±0.023.

Detailed analysis of the derived confusion matrices for each CV run using the *Correlation/Weighted* combination and $D_{Full,CV}$ revealed that on average 83% of CTRL arch shapes were correctly assigned to the CTRL group, while 13% were confused with COA and 4% were confused with ASO arch shapes (Table I). For the COA group, on average 85% were correctly assigned and the remaining 15% were confused with CTRL arch shapes. ASO arch shapes were not confused with any other shape, thus 100% were placed correctly into one ASO cluster. Notably, neither were ASO and CTRL shapes confused with high misclassification rates, nor were COA shapes always assigned correctly as we would have expected in case the clustering algorithm only took into account aortic arch size rather than shape (see section B).

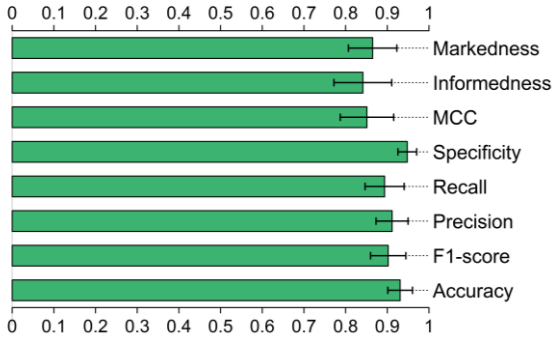

Fig. 4: Means and 95CIs over 10 CV runs of all computed clustering classification performance measures for the distance/linkage combination Correlation/Weighted, applied to the full 3D shape dataset $D_{Full,CV}$. In particular, high Specificity was achieved.

F. Subgroup Analysis – Clusters Within Clusters

Finally, clustering classification performance was assessed using the *Correlation/Weighted* distance/linkage combination and D_{Full}, including all N=60 patients. In this case, only two COA shapes (10%) were confused with CTRL arch shapes, while 100% of both CTRL and ASO arches were assigned to one respective cluster (Fig. 5a).

In order to reveal more refined shape subgroups within the three larger clusters, which would add novel information about previously unknown patterns within the pathological shape clusters, branches were cut at a lower hierarchy level. Tree branches were cut at a height of 0.72, thus forming a total of 10 subgroups with a varying number of members in each larger cluster (Fig. 5b). The CTRL group was divided into 5 smaller subgroups, the COA group into three and the ASO into two. Interestingly, the two confused COA shapes formed one distinct cluster within the CTRL group by themselves, marking them as being different from the other CTRL shapes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

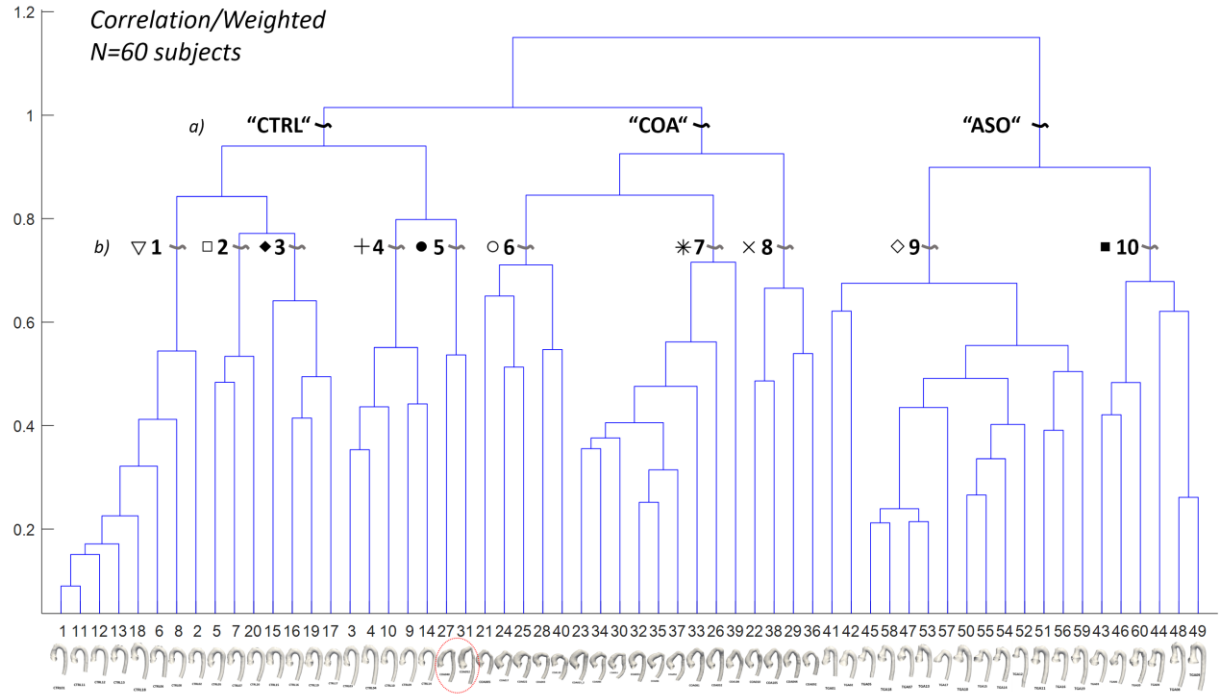

Fig. 5: Dendrogram showing shape clustering on full dataset of all patients with the Correlation/Weighted setting. The height on the Y-axis indicates the distance between subjects, computed by the linkage function. The dendrogram was automatically cut such that three large clusters emerged (a). Depending on the majority of subjects assigned to one cluster, the cluster was labelled accordingly, either CTRL, COA or ASO and confusion matrices were computed. Horizontal numbers on the X-axis represent patient identifiers: CTRL (1-20); COA (21-40) and ASO (41-60). Using the full dataset, only two COA subjects (subjects 27 and 31, marked) were misclassified as CTRL. In a second step, the dendrogram was cut at a lower level in order to reveal subgroups within the three main clusters (b). Ten subgroups were obtained as indicated. The subgroup-specific symbols are used for visualisation of subgroup affiliation in Fig. 6.

To evaluate whether the 10 subgroups related to meaningful 3D shape groups within the CTRL, COA and ASO clusters, we produced a scatter plot of the PCA shape space generated by the $\{f_{i,1}\}$ and $\{f_{i,2}\}$ associated with PCA shape modes 1 and 2 and symbol-coded the respective members of the 10 subgroups according to their subgroup affiliation (Fig. 6). This plot revealed that novel and meaningful shape subgroups within the three larger (known) shape clusters could be found,since arch shapes that were clustered together by the hierarchical clustering algorithm were also clustered closer together in terms of their 3D aortic arch shapes as described by the PCA loadings. Those findings confirmed that our pipeline can be used to detect to date unknown anatomical subgroups and patterns within pathological arch shape populations, which may prove to differ in terms of clinical outcome in future studies of larger, homogeneous patient cohorts.

TABLE I
NORMALISED CONFUSION MATRIX

Group	CTRL, predicted	COA, predicted	ASO, predicted
CTRL, actual	**83±13%**	13±13%	4±6%
COA, actual	15±10%	**85±10%**	0
ASO, actual	0	0	**100±0%**

Normalised confusion matrix for *Correlation/Weighted* distance/linkage combination. Means and 95CIs of percentage of assigned subjects from the respective groups for 10 CV runs are reported. The group that was most confused with others was the CTRL group, while ASO patients were always gathered in one cluster.

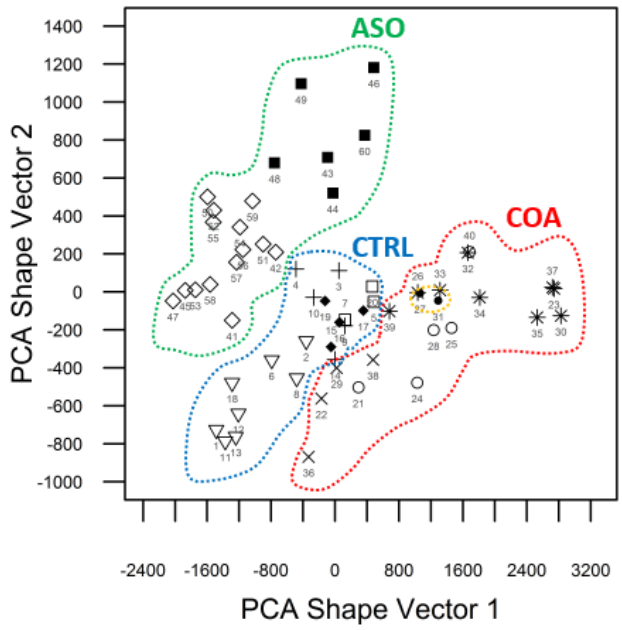

Fig. 6: Scatterplot of 3D shape space described by subject-specific PCA shape vector entries. Individual patients are symbol-coded according to subgroups obtained from cutting the dendrogram at lower levels of the hierarchy (Fig. 5b), revealing that patients with similar 3D aortic arch shape are grouped together by both the PCA analysis and the clustering algorithm. The two COA subjects misclassified as CTRL (subjects 27 and 31) are marked in orange.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2017.2655364, IEEE Transactions on Biomedical Engineering

IV. Discussion

Comparing different types of input data for the unsupervised clustering pipeline, our results showed that a preceding dimensionality reduction via PCA yielded overall lower macro *F-score*$_M$ and *MCC*$_M$ scores than the raw deformation vector data. PCA thus did not yield improved clustering classification performance, which is in accordance with previous studies [42]. Using the full deformation vector data as input, the distance/linkage combinations *Spearman/Weighted*, *Correlation/Weighted* and *Euclidean/Ward* showed overall good ability to automatically structure the bulk input data into the three clinically defined groups as measured by average *F-scores*$_M$ above 0.7 and *MCC*$_M$ scores above 0.5, following 10-fold cross-validation. This is in accordance with early observations from Lance and Williams [43] stating that the Correlation distance is suitable for comparing shapes, while the Euclidean distance is generally compatible with many clustering scenarios, probably due to being invariant under translations of the origin and under rotations of the pattern space [44]. The *Correlation* metric may here have resulted in best classification performance as it predominantly measures interrelationships between features (rather than absolute values or magnitudes) – here parameterised by shape deformation vectors of a template shape defined in a common mathematical framework. In accordance with this study, *Correlation* and *Euclidean* distance metrics have previously been found appropriate for various hierarchical clustering tasks [15], [43], [45] and so has the *Ward* linkage, specifically when clustering anatomical structures [5], [46], [47]. The *Correlation/Weighted* combination performed best with average *Specificity*$_M$, *Recall*$_M$, *Precision*$_M$ and *Accuracy*$_M$ scores around 0.9 and small confidence intervals – considerably higher than previously reported accuracies [48]. Averaged over all cross-validation runs, 17% of CTRL shapes, 15% of COA shapes and 0% of ASO shapes were misclassified. Those were lower misclassification rates than reported earlier for hierarchical clustering by Dalton (21% to 28%) [15], and Brun (13% to 48%) [14]. Applied to the full dataset of 60 patients, only two COA arch shapes that showed highly localised deformation of the transverse aortic arch were confused with CTRL shapes. This suggests that some subtle 3D arch shape features may not be taken into account sufficiently when computing inter-subject distances. This could be addressed in future studies by a weighting of local 3D shape features, depending on which section of the arch (i.e. which anatomical region) is subject of interest. As expected though, ASO arches seemed to constitute a distinctly different shape cluster, allowing for 0% of misclassification, which is a notable performance for an unsupervised and automated approach.

Furthermore, hierarchical clustering results were compared to results from PCA statistical shape analysis and found that both methods compared well in determining shape clusters and subgroups based on the deformation data. More importantly, apart from distinguishing the three clinically known groups (CTRL, COA, ASO) mostly correctly, the clustering algorithm was able to cluster together subjects with similar 3D arch shape on lower levels of the clustering tree as well. This allowed for detection of previously unknown "clusters within the cluster", i.e. novel anatomical patterns within the pathological COA and ASO clusters. While we refrained from analyzing those subgroups further due to a limited subgroup sample size, such subgroups may be discovered in future studies of larger patient cohorts via the proposed pipeline and may generate novel hypotheses of clinical relevance.

Following these results, we foresee potential application of hierarchical clustering algorithms for medical image analysis in four main areas: research, clinical, technical and commercial applications. In research, such approaches could help better understand diseases by providing a means to derive novel (anatomical, shape) biomarkers and detect yet-to-be-discovered disease patterns. This, in turn, could ultimately assist clinicians in decision-making and risk stratification, particularly in complex or rare diseases. Large, cloud-based image databases – in combination with immediate online clustering following image acquisition – could allow comparison of a newly scanned patient to individuals with the same clinical history or disease in order to detect "outliers" or similarities [49]. On the technical side, hierarchical clustering could be used for shape-retrieval systems and found clusters could be used to compute *subtemplates* or *subatlases* (i.e. representatives of a subgroup), which may improve atlas-based image segmentation of highly varying anatomy [7]. Following the overall good classification results for our unsupervised pipeline, we further assume that trained, supervised approaches would perform even better in case classification of shapes is desired. Finally, regarding commercial applications, subgroup-based subtemplate anatomical models could allow for more cost-effective "few-sizes-fit-all" rather than patient-specific approaches for device design and development, which may be particularly appealing in complex structural disease.

For such broad application of hierarchical clustering algorithms to become reality, large medical image databases are required, which leads to one of the limitations of our study – the relatively small sample size. Nevertheless, we believe this study constitutes a first step showcasing that clinically meaningful clustering of medical image data can be achieved once clustering parameters are set correctly. Future studies should focus on including more patients, different types of anatomy and on automating the pipeline further. Here, we aimed to automate data processing as much as possible. Yet, steps such as isolating the structure of interest after segmentation (here the aortic arch) were performed using manual cutting tools. This is another limitation, which may be addressed by providing segmentation atlases specifically adapted to the structure of interest. Further, some automatic segmentation results had to be edited manually due to insufficient input image quality or artefacts. With sophisticated automatic segmentation algorithms currently being on the rise [50], we foresee drastic improvement in this area in the near future. In this regard, this study reports one of the largest datasets of automatically segmented pathological structures affected by congenital heart disease and the reported Dice Similarity Coefficients could be used as reference values for further algorithm development.

V. Conclusion

In this study, we present and evaluate a medical image processing pipeline combining automatic segmentation, statistical shape modelling and unsupervised hierarchical clustering of 3D anatomical models in a cohort of healthy and pathological aortic arches post-surgical repair. By applying a specific set of distance metric and linkage function, clustering classification results yielded clinically meaningful shape clusters and subgroups – automatically derived without any prior information. To the best of our knowledge, this is the first study evaluating 3D hierarchical shape clustering performance on realistic, clinically acquired cardiovascular image data. The reported clustering classification performance and automatic segmentation scores could be used as benchmark values for future algorithm implementation and improvement.

Apart from yielding a clinically meaningful division of the data according to known clinical diagnosis, our analysis revealed novel subgroups within the known clusters, which offers the potential of providing additional information and insight into yet-to-be unveiled similarities and patterns within a disease, once an initial diagnosis has been made. Therefore, our analytical platform can be an adjunct in moving away from a case-by-case image-based diagnosis towards assessing a patient in the context of a patient population as an integral component of current Precision Medicine or "Precision Imaging" [3] strategies. Such a clinical decision support system may pave the way for moving from mere data towards information and knowledge, which could ultimately impact on improved diagnosis, risk stratification and treatment strategies.

Acknowledgment

The authors would like to thank the anonymous Reviewers for their comments that helped to improve the manuscript.

This report incorporates independent research from the National Institute for Health Research Biomedical Research Centre Funding Scheme. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

The authors gratefully acknowledge support from Fondation Leducq; Wellcome Trust (WT101957) ; Engineering and Physical Sciences Research Council (EPSRC NS/A000027/1); FP7 integrated project MD-Paedigree (partially funded by the European Commission) and Commonwealth Scholarships.

MOCHA Collaborative Group: Andrew Taylor, Alessandro Giardini, Sachin Khambadkone, Silvia Schievano, Marc de Leval and T.-Y. Hsia (Institute of Cardiovascular Science, UCL, London, UK); Edward Bove and Adam Dorfman (University of Michigan, Ann Arbor, MI, USA); G. Hamilton Baker and Anthony Hlavacek (Medical University of South Carolina, Charleston, SC, USA); Francesco Migliavacca, Giancarlo Pennati and Gabriele Dubini (Politecnico di Milano, Milan, Italy); Alison Marsden (Stanford University, Stanford, CA, USA); Irene Vignon-Clementel (INRIA,Paris,France); Richard Figliola and John McGregor (Clemson University, Clemson, SC, USA).

References

[1] H. Völzke *et al.*, "Population Imaging as Valuable Tool for Personalized Medicine," *Clinical Pharmacology & Therapeutics*, vol. 92, no. 4, pp. 422–424, Oct. 2012.

[2] P. Medrano-Gracia *et al.*, "Challenges of Cardiac Image Analysis in Large-Scale Population-Based Studies," *Curr Cardiol Rep*, vol. 17, no. 3, pp. 1–7, Feb. 2015.

[3] A. F. Frangi *et al.*, "Precision Imaging: more descriptive, predictive and integrative imaging," *Medical Image Analysis*, vol. 33, pp. 27–32, Oct. 2016.

[4] C. Petitjean *et al.*, "Right ventricle segmentation from cardiac MRI: A collation study," *Medical Image Analysis*, vol. 19, no. 1, pp. 187–202, Jan. 2015.

[5] J. J. Cerrolaza *et al.*, "Automatic multi-resolution shape modeling of multi-organ structures," *Medical Image Analysis*, vol. 25, no. 1, pp. 11–21, Oct. 2015.

[6] F. Zhao *et al.*, "Congenital Aortic Disease: 4D Magnetic Resonance Segmentation and Quantitative Analysis," *Med Image Anal*, vol. 13, no. 3, pp. 483–493, Jun. 2009.

[7] M. A. Zuluaga *et al.*, "Voxelwise atlas rating for computer assisted diagnosis: Application to congenital heart diseases of the great arteries," *Medical Image Analysis*, vol. 26, no. 1, pp. 185–194, Dec. 2015.

[8] D. Kutra *et al.*, "Automatic Multi-model-Based Segmentation of the Left Atrium in Cardiac MRI Scans," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds. Springer Berlin Heidelberg, 2012, pp. 1–8.

[9] M. B. Eisen *et al.*, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci U S A*, vol. 95, no. 25, pp. 14863–14868, Dec. 1998.

[10] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *WIREs Data Mining Knowl Discov*, vol. 2, no. 1, pp. 86–97, Jan. 2012.

[11] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[12] T. Hastie *et al.*, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009.

[13] M. Halkidi *et al.*, "On Clustering Validation Techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2–3, pp. 107–145, Dec. 2001.

[14] M. Brun *et al.*, "Model-based evaluation of clustering validation measures," *Pattern Recognition*, vol. 40, no. 3, pp. 807–824, Mar. 2007.

[15] L. Dalton *et al.*, "Clustering Algorithms: On Learning, Validation, Performance, and Applications to Genomics," *Curr Genomics*, vol. 10, no. 6, pp. 430–445, Sep. 2009.

[16] A. Srivastava *et al.*, "Statistical shape analysis: clustering, learning, and testing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 590–602, Apr. 2005.

[17] A. Dong *et al.*, "CHIMERA: Clustering of Heterogeneous Disease Effects via Distribution Matching of Imaging Patterns," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 612–621, Feb. 2016.

[18] D. Broggio *et al.*, "Comparison of organs' shapes with geometric and Zernike 3D moments," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 3, pp. 740–754, Sep. 2013.

[19] P. Ou *et al.*, "Late systemic hypertension and aortic arch geometry after successful repair of coarctation of the aorta," *European Heart Journal*, vol. 25, no. 20, pp. 1853–1859, Oct. 2004.

[20] Y. Lecompte *et al.*, "Anatomic correction of transposition of the great arteries," *J. Thorac. Cardiovasc. Surg.*, vol. 82, no. 4, pp. 629–631, Oct. 1981.

[21] J. L. Bruse *et al.*, "A Non-parametric Statistical Shape Model for Assessment of the Surgically Repaired Aortic Arch in Coarctation of the Aorta: How Normal is Abnormal?," in *O. Camara et al. (Eds.): Statistical Atlases and Computational Models of the Heart 2015*, Munich, 2016, vol. LNCS 9534, pp. 21–29.

[22] H. N. Ntsinjana *et al.*, "3D morphometric analysis of the arterial switch operation using in vivo MRI data," *Clin. Anat.*, vol. 27, no. 8, pp. 1212–1222, Nov. 2014.

[23] M. Vaillant and J. Glaunès, "Surface Matching via Currents," in *Information Processing in Medical Imaging*, G. E. Christensen and M. Sonka, Eds. Springer Berlin Heidelberg, 2005, pp. 381–392.

[24] S. Durrleman *et al.*, "Statistical models of sets of curves and surfaces based on currents," *Medical Image Analysis*, vol. 13, no. 5, pp. 793–808, Oct. 2009.

[25] M. A. Zuluaga *et al.*, "Multi-atlas Propagation Whole Heart Segmentation from MRI and CTA Using a Local Normalised Correlation Coefficient Criterion," in *Functional Imaging and Modeling of the Heart*, S. Ourselin, D. Rueckert, and N. Smith, Eds. Springer Berlin Heidelberg, 2013, pp. 174–181.

[26] P. A. Yushkevich *et al.*, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, Jul. 2006.

[27] W. Schroeder *et al.*, *Visualization Toolkit: An Object-Oriented Approach to 3D Graphics, 4th Edition*, 4th edition. Clifton Park, N.Y.: Kitware, 2006.

[28] J. Ahrens *et al.*, "ParaView: An End-User Tool for Large-Data Visualization," *The Visualization Handbook*, p. 717, 2005.

[29] L. Antiga *et al.*, "An image-based modeling framework for patient-specific computational hemodynamics," *Med Biol Eng Comput*, vol. 46, no. 11, pp. 1097–1112, Nov. 2008.

[30] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[31] J. L. Bruse *et al.*, "A statistical shape modelling framework to extract 3D shape biomarkers from medical imaging data: assessing arch morphology of repaired coarctation of the aorta," *BMC Medical Imaging*, vol. 16, p. 40, 2016.

[32] S. Durrleman *et al.*, "Morphometry of anatomical shape complexes with dense deformations and sparse parameters," *NeuroImage*, vol. 101, pp. 35–49, Nov. 2014.

[33] M. F. Beg *et al.*, "Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms," *Int J Comput Vision*, vol. 61, no. 2, pp. 139–157, Feb. 2005.

[34] M. Piccinelli *et al.*, "A Framework for Geometric Analysis of Vascular Structures: Application to Cerebral Aneurysms," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1141–1155, Aug. 2009.

[35] T. Mansi *et al.*, "A Statistical Model for Quantification and Prediction of Cardiac Remodelling: Application to Tetralogy of Fallot," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1605–1616, 2011.

[36] Joliffe, I.T., *Principal Component Analysis*, 2nd ed. Springer-Verlag New York, Inc., 2002.

[37] F. Murtagh, "A Survey of Recent Advances in Hierarchical Clustering Algorithms," *The Computer Journal*, vol. 26, no. 4, pp. 354–359, Jan. 1983.

[38] MATLAB R2011, The Mathworks, Inc., "MATLAB Help Documentation." The Mathworks, Inc., Natick MA, 2011.

[39] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[40] Powers D.M.W., "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, Dec. 2011.

[41] P. Baldi *et al.*, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, Jan. 2000.

[42] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, Jan. 2001.

[43] G. N. Lance and W. T. Williams, "A General Theory of Classificatory Sorting Strategies 1. Hierarchical Systems," *The Computer Journal*, vol. 9, no. 4, pp. 373–380, Jan. 1967.

[44] R. Dubes and A. K. Jain, "Clustering techniques: The user's dilemma," *Pattern Recognition*, vol. 8, no. 4, pp. 247–260, Oct. 1976.

[45] E. R. Dougherty *et al.*, "Inference from clustering with application to gene-expression microarrays," *J. Comput. Biol.*, vol. 9, no. 1, pp. 105–126, 2002.

[46] D. M. Boyer *et al.*, "A New Fully Automated Approach for Aligning and Comparing Shapes," *Anat. Rec.*, vol. 298, no. 1, pp. 249–276, Jan. 2015.

[47] M. Schecklmann *et al.*, "Cluster analysis for identifying sub-types of tinnitus: A positron emission tomography and voxel-based morphometry study," *Brain Research*, vol. 1485, pp. 3–9, Nov. 2012.

[48] N. S. D. Singh, "Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time."

[49] A. Tsymbal *et al.*, "Towards cloud-based image-integrated similarity search in big data," in *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014, pp. 593–596.

[50] X. Zhuang, "Challenges and Methodologies of Fully Automatic Whole Heart Segmentation: A Review," *Journal of Healthcare Engineering*, vol. 4, no. 3, pp. 371–407, 2013.