

Cardiac Population Image Quantification: Analysis of 20K Subjects in the UK Biobank

Rahman Attar^a, Marco Pereañez^a, Ali Gooya^b, Xènia Albà^c, Le Zhang^b, Stefan K. Piechnik^d,
Stefan Neubauer^d, Steffen E. Petersen^e, Alejandro F. Frangi^{a,*}

^a*Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB),
School of Computing and School of Medicine, University of Leeds, UK.*

^b*Department of Electronic and Electrical Engineering, University of Sheffield, UK.*

^c*Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB),
Universitat Pompeu Fabra, Barcelona, Spain.*

^d*Oxford Centre for Clinical Magnetic Resonance Research (OCMR),
Division of Cardiovascular Medicine, University of Oxford, John Radcliffe Hospital, UK.*

^e*Cardiovascular Medicine at the William Harvey Research Institute,
Queen Mary University of London and Barts Heart Centre, Barts Health NHS Trust, UK.*

Abstract

The analysis of large-scale population imaging datasets has the potential to improve healthcare by discovering and understanding patterns and trends of disease that translate into pre-emptive care, and improved disease management. To date, few truly large cardiac imaging studies have been available to the research community, and none with the quality and expert annotations provided by the UK Biobank (UKB) cardiac imaging project. With studies of this magnitude ($> 100,000$ subjects) a paradigm shift in the approach to image quantification is required. Before truly large datasets, single module segmentation pipelines were sufficient to analyse rather small cohorts. However, with access to large biobanks, high-throughput, quality-aware fully-automatic parsing pipelines are essential for any reliable results to be expected. In this paper we present such a pipeline, and use the UKB as case study to show its potential. The proposed pipeline performs end-to-end image processing from raw image to morphological and functional quantification while controlling input image, and output segmentation quality; all without user interaction. To the best of our knowledge, this is the first paper tackling the fully automatic 3D statistical analysis of cardiovascular shape from cardiac MR images of the UKB population study, and providing global and regional reference ranges for all key cardiovascular functional indexes, from both left and right ventricles of the heart. We validate our workflow on a reference cohort of 4620 subjects for which manual delineations, and reference functional indexes exist. Our results show significant agreement between the manually obtained reference indexes, and those automatically computed using our framework. In addition, we report the first analysis on the largest cohort of cardiac MR images 20K subjects each comprised of 50 timepoints, i.e., one million MR volumes in total.

Keywords: UK Biobank, Cardiac MR, Quality Assessment, Statistical Shape Models, Population Imaging, Fully Automatic Analysis, Cardiac Functional Indexes, Cardiac Morphological Analysis

1. Introduction

Cardiovascular diseases (CVDs) are recognised as the number one cause of death worldwide [1]. Diagnosis of cardiovascular disease is often made at

late symptomatic stages, which leads to late interventions and decreased efficacy of medical care. Early quantitative assessment of cardiac function, allowing for preventive care and early cardiovascular treatment, is therefore paramount.

Large scale population-based imaging studies of CVDs are becoming possible due to the advent of standardised robust non-invasive imaging meth-

*Corresponding author

Email address: a.frangi@leeds.ac.uk
(Alejandro F. Frangi)

ods and infrastructure for big data analysis [2]. This opens opportunities to gain new information about the development and progression of heart disease across population groups [3, 4]. Analysis and interpretation of cardiac structural and functional indexes in large scale population image data can reveal patterns and trends across population groups and, accordingly, allow insights into risk factors before CVDs develop. UKB is one of the worlds largest prospective population studies, established to investigate the determinants of disease [5]. In addition to the collection of extensive baseline questionnaire data, biological samples and physical measurements, cardiovascular magnetic resonance (CMR) is utilized to provide cardiovascular imaging-derived phenotypes [6]. CMR is part of a multi-organ, multi-modality imaging visit in 3-4 dedicated UK Biobank imaging centres that will be acquiring and store imaging data from 100,000 participants by 2022.

In terms of population sample size, experimental setup, and quality control, the most reliable reference ranges for cardiovascular structure and function found in the literature are those reported in [7], where CMR scans were manually delineated and analysed using cvi42 post-processing software (Version 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Canada). These reference values of 4620 subjects are used in this paper to validate the proposed workflow.

In addition to comparing against manual measurements, we also compare our performance against state of the art methods. In [8], the authors propose a 2D CNN-based segmentation method to analyse the UKB CMR images. Due to the discriminative nature of CNNs, segmentations lack global constraints which may be valuable or even essential for further structural analysis of muscular tissue function. In contrast, our 3D generative-based approach ensures global coherence of the cardiac anatomy and naturally lends itself to further analyses where full 3D anatomy is necessary, for instance, in mechanical simulations.

In this paper, we present a novel fully automatic 3D image parsing workflow with quality control modules to analyse CMR images in the UKB and corroborate their validity by comparing to their manual counterpart, and state of the art methods [8]. The proposed workflow is capable of segmenting the cardiac ventricles and generating global and regional clinical reference ranges that are statistically comparable to those obtained by human ob-

servers. The main contribution of this paper is in its clinical impact, resulting from the analysis of left ventricle (LV) and right ventricle (RV) of the heart, as well as the extraction of key cardiac functional indexes from a large CMR dataset of 20K subjects each comprised of 50 timepoints, i.e. one million CMR scans.

2. Methodology

There are four main aspects to our image parsing pipeline: 1) pre-processing, 2) quality awareness, 3) segmentation, and 4) quantification. Figure 1 illustrates the proposed workflow. In the following sections we present every step of our pipeline in detail.

2.1. Pre-processing

The UKB study conducts detailed MRI image scans of all vital organs in the body using specialised imaging protocols. The data is stored in servers due to its size, but most importantly, for ready-access to HPC resources. The first step in processing the specific subset of cardiac data we are interested in, i.e., ventricular function CINE studies, is the localisation of all relevant series, such as, short axis (SAX) and long axis (LAX) 2, 3 and 4-chamber images. Additionally, parallelisation infrastructure is necessary for efficient use of HPC and cloud resources.

2.1.1. Data Organisation

A data organisation module was developed to automatically identify relevant image series from raw DICOM data, and organise these data hierarchically per patient. This pre-processing step provides database searchability and interpretability for the user. Descriptive and unique file names were used to reflect the contents of files and uniquely identify images enabling accessibility and discovery. The folder structure of every subject in the dataset contains image information for 50 cardiac cycle timepoints. Every timepoint contains Short Axis (SAX) and Long Axis (LAX) subfolders. The SAX folder contains slices of short axis view and the LAX folders contain 2, 3 and 4-chamber views of the heart.

2.1.2. Parallelisation

To create a modular workflow and run it in parallel for multiple subjects, a workflow manager software package is required. Such manager provides

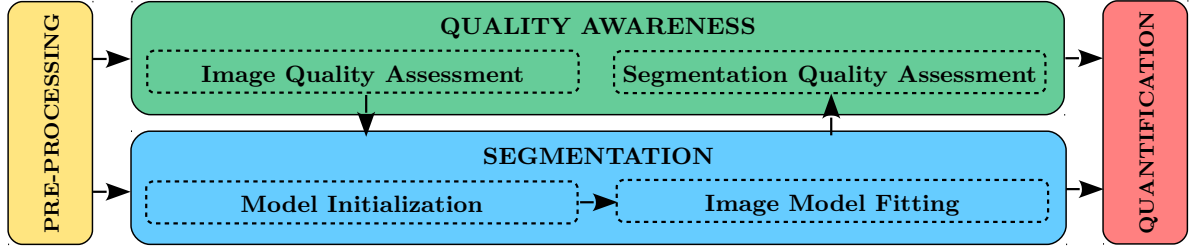


Figure 1: Fully automatic image parsing workflow for large scale analysis of cardiac ventricles.

the parallelisation infrastructure, software package inter-operability, and workflow monitoring tools. In this work, the Nipype software package [9] within the MULTIX platform¹ is used. Nipype allows us to integrate software packages, written in different computer languages, within a single workflow, and parallelise the tasks using cloud computing resources. Here, we use Amazon² high-performance processors and S3 storage services to store and analyse the CMR database.

2.2. Quality Awareness (QA)

To ensure the reliability of any large throughput image analysis task, we believe at least two quality control modules are necessary. One such module should be dedicated to assessing the quality of input images to the pipeline. The other module, should assess the reliability of outputs; in our case, 3D segmentations. Both this modules are included within our processing pipeline and we denote them "Image Quality Assessment" (IQA), and "Segmentation Quality Assessment" (SQA) modules, respectively.

2.2.1. Image Quality Assessment (IQA)

Despite careful and strict imaging guidelines, in any large-scale study, a certain portion of the data can be expected to fall outside the guidelines. To ensure that the quality of the collected data optimises the accuracy of segmentation results, an IQA module has been included to detect abnormal images whose analysis would otherwise impair any aggregated statistics over the cohort. Since the lack of basal and/or apical slices is the most common problem affecting image quality in CMR images, and their absence has a major impact on the accuracy of quantified parameters [10], the IQA module

detects missing basal/apical slices in the input SAX volumes. Every top and bottom short-axis view slice from input SAX volumes is run through an independent CNN classifier that determines presence/absence of the basal and apical slices. The details of the architecture used can be found in [11].

2.2.2. Segmentation Quality Assessment (SQA)

The large anatomical variability found in population studies [12], as well as, aspects of image quality not captured by IQA, can cause image segmentations to fail. A self-verification mechanism is therefore necessary to automatically detect incorrect segmentations. These images can then either be re-processed with adjusted parameters, or discarded from latter statistical analysis. In our pipeline we incorporate the SQA proposed in [13]. The SQA module uses Random Forest classifiers trained on intensity features associated to blood pool and myocardium, and is able to detect successful segmentations.

2.3. Segmentation

SAX and LAX images are used to make a first estimate of the position and orientation of the ventricles. With this information we then initialise the segmentation of the cardiac structure following a Sparse Active Shape Model (SPASM) approach. SPASM is used to segment the full cardiac cycle, and retrospectively determine the end-diastolic (ED) and end-systolic (ES) phases of the cycle based on maximum and minimum volume computation. Prior to segmentation, we performed grid optimisation of those parameters having the highest impact on segmentation performance (see Sec. 2.3.3).

2.3.1. Model Initialisation

To initialise the model automatically, the method proposed in [13] has been used with a further step

¹<https://multi-x.org>

²<https://aws.amazon.com>

to improve biventricular model initialisation. First, the location of the LV is determined by a rough estimation of the intersection of slices from SAX and LAX views. Then, a random forest regressor trained with two complementary feature descriptors (i.e. the Histogram of Oriented Gradients and Gabor Filters) is used to predict correct landmark positions for the LV. We extended this to take into account image features corresponding to the RV, and improve the initial estimate for the biventricular heart. These landmarks are used to estimate the pose parameters that place a mean shape model near the heart. These parameters are used to initialise the first image volume in the set of images for the cardiac cycle (50 timepoints), subsequent timepoints are automatically initialised with the shape model fitted to the previous timepoint image.

2.3.2. Image Model Fitting

The cardiac LV and RV segmentation are performed using the SPASM segmentation method [14], which improves on Active Shape Models (ASM) [15] by addressing the sparsity found in imaging modalities such as CMR where image information is sparsely distributed in the image volume. The main components of the SPASM are a Point Distribution Model (PDM), an Intensity Appearance Model (IAM), and a model matching algorithm.

The PDM encodes the mean and variance of the endo and epicardial shapes of the LV and the endocardial shape of the RV. The PDM is constructed during training using Principal Component Analysis (PCA) on a set of generalized procrustes-aligned shapes, preserving 98% variance.

Let us assume a training set of M shapes each described by N points in \mathcal{R}^3 , $\mathbf{x}_j^i = (\mathbf{x}_j^i, \mathbf{y}_j^i, \mathbf{z}_j^i)$ with $i = 1, \dots, M$ and $j = 1, \dots, N$.

Let $\mathbf{s}_i = (\mathbf{x}_1^i, \mathbf{y}_1^i, \mathbf{z}_1^i, \dots, \mathbf{x}_N^i, \mathbf{y}_N^i, \mathbf{z}_N^i)^T$ be the i -th vector representing the shape of the i -th endocardial and epicardial surfaces of LV and the endocardial surface of RV. Finally, let $\mathbf{S} = [\mathbf{s}^1, \dots, \mathbf{s}^M]$ be the set of all training shapes in matrix form. All nuisance pose parameters (e.g. translation, rotation and scaling) have been removed from S using generalized procrustes analysis. The shape class mean and covariance of \mathbf{S} is as follows:

$$\bar{\mathbf{s}} = \frac{1}{M} \sum_{i=1}^M \mathbf{s}_i \quad (1)$$

$$\mathbf{C} = \frac{1}{M-1} \sum_{i=1}^M (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^T \quad (2)$$

The shape covariance is represented in a low-dimensional space or PCA of the shape. This produces l eigenvectors $\Phi = [\varphi_1 \varphi_2 \dots \varphi_l]$, and corresponding eigenvalues $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_l)$ of the covariance matrix computed via Singular Value Decomposition (SVD). Hence, assuming the shape class follows a multi-dimensional Gaussian probability distribution, any shape in the shape class can be approximated from the following linear generative model:

$$\mathbf{s} \approx \bar{\mathbf{s}} + \Phi \mathbf{b} \quad (3)$$

where \mathbf{b} are shape parameters restricted to $|\mathbf{b}_i| \leq \beta \sqrt{\lambda_i}$ typically set to $\beta = 3$ to capture 99.7% of shape variability. The shape parameters of \mathbf{s} can be estimated by

$$\mathbf{b} = \Phi^T (\mathbf{s} - \bar{\mathbf{s}}). \quad (4)$$

The entries of \mathbf{b} are the projection coefficients of mean-centred shapes $(\mathbf{s} - \bar{\mathbf{s}})$ along the columns of Φ .

For each landmark in \mathbf{s} , we build an IAM based on the intensity information across all corresponding landmarks in all training shapes \mathbf{s}_i . IAMs capture the local intensity distribution along cardiac boundaries. We proceed by sampling 1D intensity profiles normal to the myocardial boundaries. Each profile has a length size $m = 15$ pixels. For the i -th landmark, the mean intensity profile, $\bar{\mathbf{g}}_i$, and the corresponding image intensity covariance, \mathbf{S}_{g_i} , are estimated. During image segmentation, the intersections of the current shape model instance with all image planes defines a stack of 2D contours in \mathcal{R}^3 . The algorithm proceeds by searching for the best-matching intensity profile location along the normal to the contours and over the imaging planes for each landmark. To derive the best-matching position, or candidate point, \mathbf{y}_i for each landmark, we minimize the Mahalanobis distance between a profile sampled at the candidate position, i.e. $\mathbf{g}_i(\mathbf{y}_i)$, and the corresponding model, $\{\bar{\mathbf{g}}_i, \mathbf{S}_{g_i}\}$, as

$$\mathbf{y}_i^o = \arg \min_{\mathbf{y}_i} ((\mathbf{g}(\mathbf{y}_i) - \bar{\mathbf{g}}_i)^T \mathbf{S}_{g_i}^{-1} (\mathbf{g}(\mathbf{y}_i) - \bar{\mathbf{g}}_i)). \quad (5)$$

The search for candidate positions takes place normally to the LV myocardial and RV endocardial contours and over planes corresponding to CMR

imaging planes. However, the candidate positions themselves lay on 3D space and the vector between the current and candidate position of each landmark can be interpreted as a landmark displacement force. SPASM propagates these forces over neighbouring nodes weighted by the geodesic distance between the search position and neighbour nodes as

$$\mathbf{w}(p, q) = \exp\left\{-\frac{\|p - q\|^2}{2\sigma^2}\right\} \quad (6)$$

where $\|p - q\|^2$ is the geodesic distance between points p and q , and σ is the width of the Gaussian kernel that reflects the degree of sparsity of the data.

2.3.3. Parameter Optimisation

SPASM segmentation is affected by four main parameters. We ran an exhaustive grid optimisation scheme to determine the best parameter combination. The parameters tested were:

1. Freedom of PDM parameters measured in standard deviations from the mean ($\sigma = 2, 2.5, 3$).
2. Length of the image sampling profile used during image feature search ($l = 5, 10, 15$).
3. The standard deviation of a Gaussian kernel around shape points influencing the motion of points during image matching ($k = 5, 7, 9$).
4. Using only SAX images, or both SAX and LAX during segmentation ($v = SAX, ALL$).

We executed the segmentation algorithm with 54 different parameter combinations ($3 \times 3 \times 3 \times 2 = 54$), as shown in Table 1, on 50 randomly-selected subjects that had already been manually delineated by clinicians. Then we computed segmentation accuracy using the following metrics: Dice Similarity Coefficient (DSC), Mean Contour Distance (MCD) and Hausdorff Distance (HD). Refer to Equations 7, 8 and 9 in Section 3.1 for definitions of the metrics. Figure 2 shows each of the 54 executions for the three different metrics in boxplots. The set of parameters that yielded best results jointly for the three metrics was "Test 04" indicating $\sigma = 2$, $l = 5$, $k = 7$ and $v = ALL$. We therefore used this parameter set for all segmentations in our experiments.

2.4. Quantification

We computed a thorough set of functional parameters based on blood-pool and myocardial volumes. To reproduce those reference ranges reported

Table 1: List of 54 different sets of segmentation parameters used for parameter optimisation.

Test	σ	l	k	v	Test	σ	l	k	v
01	2	5	5	SAX	28	2.5	10	7	ALL
02	2	5	5	ALL	29	2.5	10	9	SAX
03	2	5	7	SAX	30	2.5	10	9	ALL
04	2	5	7	ALL	31	2.5	15	5	SAX
05	2	5	9	SAX	32	2.5	15	5	ALL
06	2	5	9	ALL	33	2.5	15	7	SAX
07	2	10	5	SAX	34	2.5	15	7	ALL
08	2	10	5	ALL	35	2.5	15	9	SAX
09	2	10	7	SAX	36	2.5	15	9	ALL
10	2	10	7	ALL	37	3	5	5	SAX
11	2	10	9	SAX	38	3	5	5	ALL
12	2	10	9	ALL	39	3	5	7	SAX
13	2	15	5	SAX	40	3	5	7	ALL
14	2	15	5	ALL	41	3	5	9	SAX
15	2	15	7	SAX	42	3	5	9	ALL
16	2	15	7	ALL	43	3	10	5	SAX
17	2	15	9	SAX	44	3	10	5	ALL
18	2	15	9	ALL	45	3	10	7	SAX
19	2.5	5	5	SAX	46	3	10	7	ALL
20	2.5	5	5	ALL	47	3	10	9	SAX
21	2.5	5	7	SAX	48	3	10	9	ALL
22	2.5	5	7	ALL	49	3	15	5	SAX
23	2.5	5	9	SAX	50	3	15	5	ALL
24	2.5	5	9	ALL	51	3	15	7	SAX
25	2.5	10	5	SAX	52	3	15	7	ALL
26	2.5	10	5	ALL	53	3	15	9	SAX
27	2.5	10	7	SAX	54	3	15	9	ALL

in [7], our quantification module performs volume computations using Simpson's rule. The principle underlying this method is that total volume can be approximated by the summation of stacks of elliptical disks.

We computed both global and regional morphological and functional indexes. Global indices include chamber volumes, stroke volume, ejection fraction and myocardial mass. Regional or local indices include myocardial wall thickness, motion and thickening based on the AHA-17 [16] cardiac subdivision scheme.

Global assessment of cardiac function relies on the following volumetric measurements [17]:

- End-Diastolic Volume (EDV): the volume of blood in the right or left ventricle at end load or filling, or the amount of blood in the ventricle just before systole.
- End-Systolic Volume (ESV): the volume of blood in the right or left ventricle at the end of contraction. This is the lowest volume of blood in the ventricle at any point in the cardiac cycle.

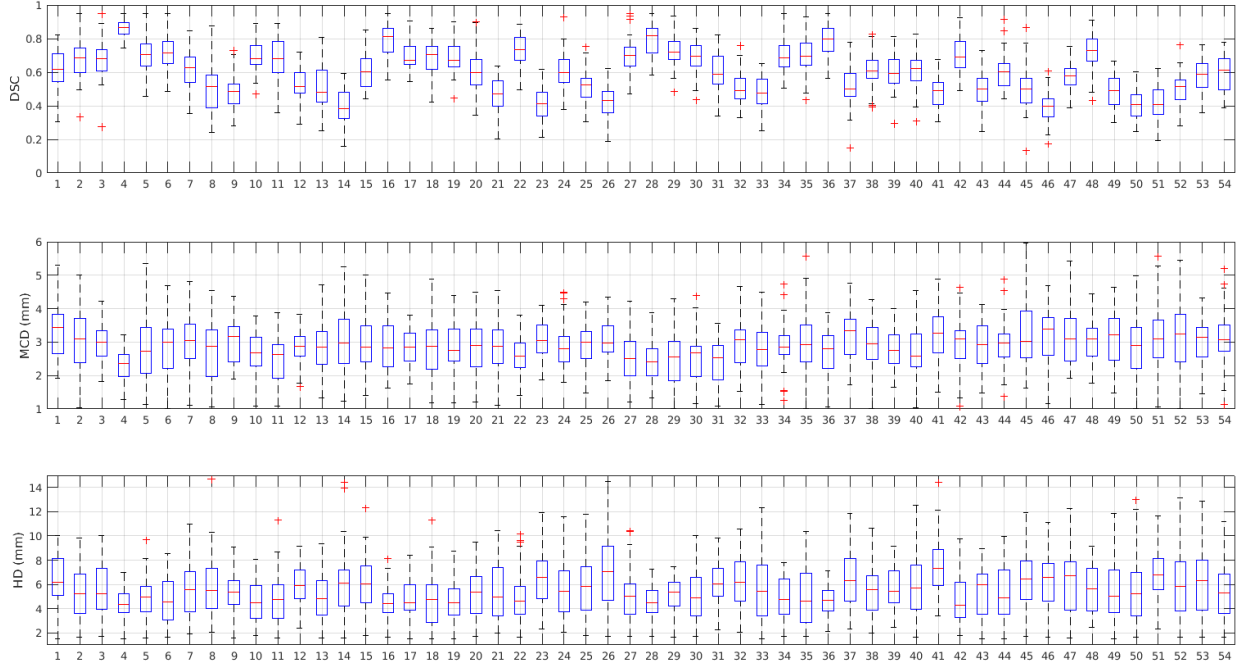


Figure 2: Parameter optimisation of the segmentation algorithm. The set of parameters that yielded best results jointly for DSC, MCD and HD metrics was experiment number "4".

- **Stroke Volume (SV):** the volume of blood pumped from the ventricle per beat. It is obtained by subtracting ESV from EDV for a given ventricle. The term stroke volume can be applied to either of the two ventricles of the heart.
- **Ejection Fraction (EF):** the fraction of blood ejected from a ventricle of the heart with each heartbeat showing the pumping efficiency of the heart. EF is calculated by dividing the SV by EDV. The EF of the left heart, known as the left ventricular ejection fraction (LVEF), is a measure of the efficiency of pumping into the body's systemic circulation. The EF of the right heart, or right ventricular ejection fraction (RVEF), is a measure of the efficiency of pumping into the pulmonary circulation (the lungs).
- **Left Ventricular Mass (LVM):** This is mainly determined by two factors: chamber volume, and wall thickness. To compute LVM, there are two main assumptions: 1) the inter ventricular septum is assumed to be part of the LV, and 2) the volume of the myocardium is equal to the total volume contained within the

epicardial borders of the ventricle minus the chamber volume. Then LVM is obtained by multiplying the volume by the density of the muscle tissue (1.05 g/cm^3).

Regional assessment of cardiac function relies on the following indexes obtained from the LV myocardium shapes and computed locally based on the AHA 17-segment model:

- **LV Wall Thickness** is the thickness of the myocardium at any timepoint of the cardiac cycle, however, it is typically measured at ED and ES. Wall thickness may be used to quantify regional dysfunction, as seen for instance, in myocardial ischemia or after myocardial infarction. The most widely employed method for Wall Thickness computation is the *centre-line method* [18].
- **LV Wall Thickening** is the thickness differential typically measured between ED and ES. Papillary muscles and trabecular tissues are usually excluded.
- **LV Wall Motion** measures the motion of the myocardial wall between ED and ES phases of the cardiac cycle.

We present and compare all above-mentioned global and regional clinical indexes obtained through manual and automatic segmentation in Section 3.

3. Experiments and Results

We evaluate the performance of the proposed automatic workflow in two ways: 1) using common metrics for segmentation accuracy assessment i.e. Dice Similarity Coefficient (DSC), Mean Contour Distance (MCD) and Hausdorff distance (HD) and comparing them against the ground truth obtained through manual delineation from clinicians; and 2) using clinical cardiac biventricular functional indexes derived from manual and automatic segmentations such as EDV, ESV and LVM.

We also compare our results with those reported in [8]. Table 2 shows the data used for training, testing and evaluating the workflow. Out of 4,870 available subjects with manual segmentations, 500 shapes (250 ED and 250 ES) from 250 random subjects were used for PDM training. 1000 image volumes (500 ED and 500 ES) from 500 random subjects from the MESA dataset [19] were used for IAM training. The rest of subjects (4,620) in the UKB with manual delineations were used as test datasets to evaluate the performance of the proposed automatic approach (A2). To compare with [8], denoted A1, we use exactly the same training/testing datasets and report the results as A3 in the tables. As an additional assessment, quantitative evaluation of human performance i.e. the inter-observer variability is measured among the segmentations done manually by different clinical experts. A set of 50 subjects was randomly selected and each subject was analysed by three expert observers (O1, O2, O3) independently. We compare the result of segmentations on the same set of subjects to show automatic vs. human performance, and also the performance of our workflow on a large dataset.

Input images and output segmentation contours were automatically quality controlled to ensure that input image volumes had full coverage of the heart, i.e., include both basal and apical slices, and to verify the quality of output segmentations. Since the aim of the results presented in Section 3.1 is the evaluation of segmentation accuracy, all segmentation results (including outliers) were included in the statistics. In contrast, those results presented in Section 3.2 are based only on good quality images and segmentations, i.e. excluding those deemed

suboptimal by SQA and/or not full coverage by IQA.

To provide a visual sense of three different qualities (Best, Average and Worst) of the LV and RV segmentations on short-axis images are shown on Figure 3. The figure shows that automatic segmentation agrees well with manual segmentation both at ED and ES and at different slice locations (apical, mid and basal regions).

3.1. Segmentation Accuracy

To quantify segmentation accuracy, we use the following metrics. The DSC evaluates the overlap between automatic segmentation \mathbf{A} and manual segmentation \mathbf{M} and it is defined as:

$$DSC = \frac{2|\mathbf{A} \cap \mathbf{M}|}{|\mathbf{A}| + |\mathbf{M}|} \quad (7)$$

DSC is defined between 0 and 1, with higher DSC indicating better match between two segmentations. The MCD and HD evaluate the mean and the maximum distance respectively between the segmentation contours $\partial\mathbf{A}$ and $\partial\mathbf{M}$. They are defined as

$$MCD = \frac{1}{2|\partial\mathbf{A}|} \sum_{p \in \partial\mathbf{A}} d(p, \partial\mathbf{M}) + \sum_{q \in \partial\mathbf{M}} d(q, \partial\mathbf{A}) \quad (8)$$

$$HD = \max(\max_{p \in \partial\mathbf{A}} d(p, \partial\mathbf{M}), \max_{q \in \partial\mathbf{M}} d(q, \partial\mathbf{A})) \quad (9)$$

where $d(p, \partial)$ denotes the minimal distance from point p to contour ∂ . The lower the distance metric, the better the agreement.

Table 3 reports DSC, MCD, and HD between automatic and manual segmentation, evaluated on test sets of 50, 600, and 4620 subjects which have not been seen before by the PDM and IAM. The set of 50 subjects is the same set that has been used for the evaluation of inter-observer variability. The set of 600 subjects is the same set used as test set in [8] where a deep learning approach was utilised for the segmentation. The set of 4620 subjects are all UKB cases with manual delineations that have not been used for shape and appearance model training.

In Table 3 the mean and standard deviations of DSC for the LV_{endo} , LV_{myo} and RV_{endo} for $n = 4620$ test dataset are 0.93 ± 0.05 , 0.87 ± 0.05 , and 0.87 ± 0.07 , respectively, which indicates excellent agreement between manual delineations and automatic segmentations. It can be seen that DSC

Table 2: Datasets used for training and testing the methods on this paper.

Name	Method	Training/Tuning Data	Test Data
A1	Proposed Method in [8]	4275 subjects from UKB	1) 600 subjects from UKB 2) 50 subjects from UKB
A2	Proposed Workflow	PDM: 250 subjects from UKB IAM: 1000 subjects from MESA	1) 4,620 subjects from UKB 2) 600 subjects from UKB 3) 50 subjects from UKB
A3	Proposed Workflow	PDM: 4275 subjects from UKB IAM: 4275 subjects from UKB	600 subjects from UKB
O1-O3	Manual Segmentation by 3 different human observer	NA	50 subjects from UKB

Table 3: Segmentation results on the testing cases in terms of DSC, MCD and HD comparing manual, automatic methods, as well as human observers. M: ground truth provided by manual segmentation [7]. LV_{endo}: LV endocardium. LV_{myo}: LV Myocardium, RV_{endo}: RV endocardium. Values indicate mean \pm standard deviation.

(a) DSC

	O1 vs O2 (n=50)	O2 vs O3 (n=50)	O3 vs O1 (n=50)	A1 vs M (n=50)	A2 vs M (n=50)	A1 vs M (n=600)	A2 vs M (n=600)	A3 vs M (n=600)	A2 vs M (n=4620)
LV _{endo}	0.94 \pm 0.04	0.92 \pm 0.04	0.93 \pm 0.04	0.94 \pm 0.04	0.93 \pm 0.03	0.94 \pm 0.04	0.93 \pm 0.05	0.94 \pm 0.04	0.93 \pm 0.05
LV _{myo}	0.88 \pm 0.02	0.87 \pm 0.03	0.88 \pm 0.02	0.87 \pm 0.03	0.88 \pm 0.03	0.88 \pm 0.03	0.87 \pm 0.04	0.87 \pm 0.03	0.87 \pm 0.05
RV _{endo}	0.87 \pm 0.06	0.88 \pm 0.05	0.89 \pm 0.05	0.86 \pm 0.07	0.87 \pm 0.06	0.90 \pm 0.05	0.88 \pm 0.06	0.89 \pm 0.05	0.87 \pm 0.07

(b) MCD (mm)

	O1 vs O2 (n=50)	O2 vs O3 (n=50)	O3 vs O1 (n=50)	A1 vs M (n=50)	A2 vs M (n=50)	A1 vs M (n=600)	A2 vs M (n=600)	A3 vs M (n=600)	A2 vs M (n=4620)
LV _{endo}	1.00 \pm 0.25	1.30 \pm 0.37	1.21 \pm 0.48	1.08 \pm 0.30	1.28 \pm 0.39	1.04 \pm 0.35	1.21 \pm 0.36	1.06 \pm 0.35	1.18 \pm 0.41
LV _{myo}	1.16 \pm 0.34	1.19 \pm 0.25	1.21 \pm 0.36	1.18 \pm 0.31	1.20 \pm 0.34	1.14 \pm 0.40	1.23 \pm 0.48	1.13 \pm 0.35	1.23 \pm 0.50
RV _{endo}	2.00 \pm 0.79	1.78 \pm 0.45	1.87 \pm 0.74	2.20 \pm 0.92	1.79 \pm 0.80	1.78 \pm 0.70	1.80 \pm 0.80	1.74 \pm 0.61	1.80 \pm 0.69

(c) HD (mm)

	O1 vs O2 (n=50)	O2 vs O3 (n=50)	O3 vs O1 (n=50)	A1 vs M (n=50)	A2 vs M (n=50)	A1 vs M (n=600)	A2 vs M (n=600)	A3 vs M (n=600)	A2 vs M (n=4620)
LV _{endo}	2.84 \pm 0.70	3.31 \pm 0.90	3.25 \pm 0.96	3.46 \pm 1.05	3.21 \pm 0.97	3.16 \pm 0.98	3.29 \pm 1.04	3.15 \pm 0.96	3.44 \pm 1.08
LV _{myo}	3.70 \pm 1.16	3.82 \pm 1.07	3.76 \pm 1.21	4.06 \pm 1.16	3.91 \pm 1.20	3.92 \pm 1.37	3.97 \pm 1.43	3.90 \pm 1.29	3.98 \pm 1.49
RV _{endo}	7.56 \pm 5.51	7.35 \pm 2.19	7.14 \pm 2.20	9.02 \pm 3.54	7.41 \pm 4.11	7.25 \pm 2.70	7.54 \pm 3.20	7.21 \pm 2.62	7.84 \pm 3.19

for the LV_{myo} and RV_{endo} is lower than that of the LV_{endo}. One reason DSC values for the LV_{myo} are lower is its larger perimeter (endo and epicardial edge), which causes equal overlap shifts to have a greater impact compared to LV_{endo} and RV_{endo}. Also segmentation of the RV is in general more challenging due to lower SNR and/or sub-pixel thinness of the RV myocardium. Additionally, the presence of trabeculations in the cavity with signal intensities similar to that of the myocardium, the more complex crescent shape of the RV, which varies from the base to the apex, difficulty in segmenting the apical image slices, and considerable variability in shape and intensity of the chamber among subjects, notably in pathological cases, are reasons that the RV is a more challenging structure to segment than the LV.

The MCD is 1.18 ± 0.41 mm for the LV_{endo},

1.23 ± 0.50 mm for the LV_{myo} and 1.80 ± 0.69 mm for the RV_{endo}, all of which are smaller than the in-plane pixel spacing range of 1.8 mm to 2.3 mm. The Hausdorff distance are 3.44 ± 1.08 mm, 3.98 ± 1.49 mm, and 7.84 ± 3.19 mm for LV_{endo}, LV_{myo} and RV_{endo}, respectively. Although HD is larger than the in-plane pixel spacing but still within an acceptable range compared with the distance range between different observers.

When comparing our method with A1, there is a notable difference in performance between A2 (relatively small training set), and A3 (same training as A1). On Table 3 a slight improvement of the mean and standard deviation can be observed and particularly for MCD. Nevertheless, improvements become more apparent when looking at Figure 4 where the number of outlying subjects is drastically reduced for A3 when compared both to A1 and A2.

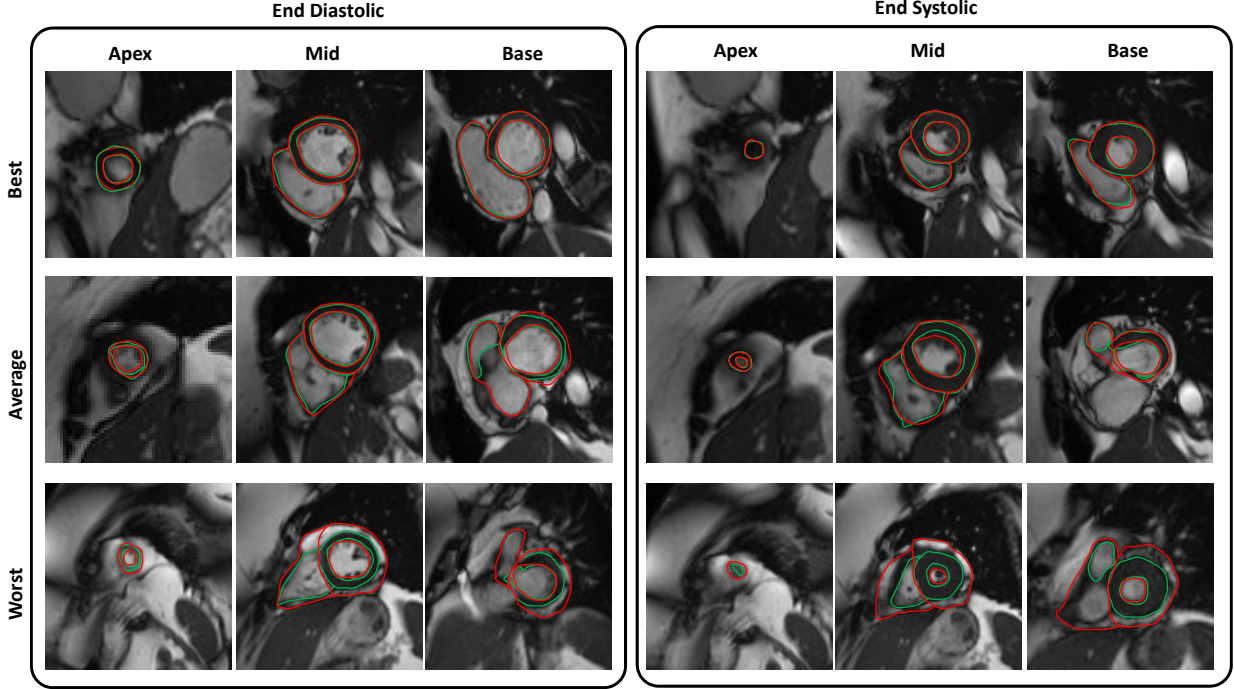


Figure 3: Examples of segmentation results at ED and ES phases: three types of the automatic segmentation contours versus manual contours. Our method is marked as red and the ground truth as green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Though the overall mean and standard deviation remains slightly better for A1. Figure 4 shows that A3 is more robust as it reduces the number and deviation of outlying results.

It is clear that the performance of A2 is in a good agreement with the ground truth and comparable to that of A1. Additionally, we looked into LV myocardium segmentation accuracy in detail based on the AHA 17-segment model [16] to report the local segmentation accuracy in terms of DSC, MCD and HD between the manual segmentation and automatic approaches, namely A1 and A2 on the 600 test set. The local segmentation accuracy is reported in Table 4. The tables shows that A1 and A2 consistently perform better in mid ventricular and apical slices, respectively. However, in base slices the performance of A1 and A2 is complementary.

Note that when comparing the performance of A2 vs. A1 ($n = 600$) in Table 3, A1 yields slightly globally better results than A2, however, upon breakdown of the results into cardiac regions, i.e., basal, mid and apical, (See Table 4) we see our method (A2) consistently outperforms A1 for all metrics in the apical region (AHA segments 13-17). Possible

reasons for this could be: 1) the inability of CNNs to capture small features in the image, and 2) the inherent ability of PDMs to infer missing or noisy image data.

Figure 4 shows the boxplots obtained from the Table 3 to demonstrate the distribution of the metrics used to assess the segmentation accuracy. The accuracy of the segmentations shows that our method (A2 and A3) can achieve error ranges observed between different human raters. This shows that our workflow performs with human-like reliability, and can fully automatically segment large scale datasets where manual inputs are infeasible.

3.2. Estimation of Cardiac Function Indexes

We evaluate the accuracy of cardiac function indexes derived from automatic segmentation using gold standard reference ranges derived from manual segmentations. Additionally we report the analysis of all available CMR images from the UKB which to date is 20K subjects. We calculate: 1) *global indexes* such as: the LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV), LV Stroke Volume (LVSV), LV Ejection-Fraction

Table 4: Regional segmentation accuracy of LV myocardium based on AHA 17-segment model on 600 subjects. Values indicate mean \pm standard deviation.

		DSC		MCD		HD	
		A1 vs M	A2 vs M	A1 vs M	A2 vs M	A1 vs M	A2 vs M
Basal	1	0.82 ± 0.03	0.84 ± 0.02	0.97 ± 0.7	0.95 ± 0.37	3.52 ± 1.00	2.29 ± 1.71
	2	0.86 ± 0.03	0.82 ± 0.03	1.10 ± 0.48	1.28 ± 0.36	2.86 ± 1.11	3.30 ± 0.92
	3	0.85 ± 0.04	0.83 ± 0.03	1.01 ± 0.36	1.10 ± 0.39	3.86 ± 1.04	2.31 ± 0.95
	4	0.85 ± 0.01	0.83 ± 0.02	0.82 ± 0.36	0.93 ± 0.34	3.59 ± 1.49	2.86 ± 1.23
	5	0.83 ± 0.03	0.85 ± 0.01	1.10 ± 0.34	1.16 ± 0.44	2.94 ± 1.41	3.12 ± 1.16
	6	0.86 ± 0.01	0.85 ± 0.03	1.07 ± 0.45	1.01 ± 0.42	3.45 ± 1.28	3.28 ± 1.02
Mid	7	0.90 ± 0.02	0.86 ± 0.03	0.88 ± 0.37	0.98 ± 0.43	2.06 ± 1.31	3.68 ± 1.24
	8	0.91 ± 0.03	0.86 ± 0.02	1.14 ± 0.42	1.20 ± 0.45	3.42 ± 1.26	3.72 ± 1.33
	9	0.89 ± 0.03	0.87 ± 0.02	1.04 ± 0.31	1.08 ± 0.38	2.63 ± 1.30	3.80 ± 0.93
	10	0.88 ± 0.02	0.87 ± 0.02	1.34 ± 0.37	1.49 ± 0.36	2.76 ± 1.22	3.88 ± 1.09
	11	0.90 ± 0.03	0.88 ± 0.02	1.16 ± 0.43	1.24 ± 0.41	2.50 ± 1.13	3.52 ± 0.90
	12	0.90 ± 0.04	0.88 ± 0.03	1.03 ± 0.33	1.06 ± 0.44	3.00 ± 1.27	3.65 ± 1.19
Apical	13	0.86 ± 0.02	0.88 ± 0.02	1.39 ± 0.40	1.24 ± 0.43	5.60 ± 1.10	4.20 ± 1.21
	14	0.87 ± 0.02	0.89 ± 0.02	1.58 ± 0.42	1.53 ± 0.43	5.16 ± 1.09	4.26 ± 1.32
	15	0.88 ± 0.02	0.90 ± 0.02	1.76 ± 0.48	1.56 ± 0.46	5.60 ± 1.11	4.31 ± 0.95
	16	0.89 ± 0.03	0.91 ± 0.02	1.83 ± 0.43	1.59 ± 0.40	5.64 ± 1.15	4.71 ± 1.24
Apex	17	0.91 ± 0.03	0.93 ± 0.03	2.00 ± 0.44	1.83 ± 0.45	5.40 ± 1.17	4.81 ± 1.14

(LVEF), LV myocardial mass (LVM), RV end-diastolic volume (RVEDV) and end-systolic volume (RVESV), RV Stroke Volume (RVSV) and RV Ejection-Fraction (RVEF); and 2) *regional indexes* such as: the myocardium wall thickness, thickening and motion, all from automated segmentation and compare them to measurements from manual segmentation.

Here, we report the clinical indexes obtained from automatic segmentation of the subjects that have passed the IQA and SQA module successfully. Table 5 shows the number of subjects that went through the analysis. For instance, out of 4620 subjects, 4430 subjects were deemed of good quality after IQA and SQA analysis, (IQA detected 145 subjects, SQA detected 105 subjects; 60 subjects were common in both lists). Thus, in total, 190 subjects were automatically removed before the analysis.

Table 6 shows the main cardiac clinical indexes. The first two columns represent the ventricular parameters of healthy population obtained through automatic and manual segmentations. It is clear that there is a good agreement between the two methods for computing cardiac function indexes. Likewise, the computed clinical indexes for a large cohort of 4620 subjects are correlated well with the ground truth values, as shown in columns three

and four. However, we notice that although the mean and standard deviation of the RV indexes for healthy population of 800 subjects is in a good agreement, the mean and standard deviation of RV indexes is slightly different for the population of 4620 subjects compared with the ground truth. This correlates with the larger inter-observer variability shown on Table 3, which is at least in part due to lower SNR and thinness of the RV myocardium vis-a-vis the LV [20].

Table 7 shows the mean absolute and relative differences between the automatic and manual measurements, and between measurements computed by different expert observers and also by the built-in software of the scanner device (inlineVF D13A). The absolute and relative differences for two subsets of 50 and 600 subjects match well and are within the error range of three different expert observers. Similarly, although the range of difference over the cohort of 4620 subjects are not directly comparable with a test set of 50, the difference range still is either within that range or very close to the difference range obtained by the different expert observers. A1, A2, and A3 perform by far better than the automatic segmentation obtained from the inlineVF D13A software; these data have been retrieved for every subject in the main UKB database.

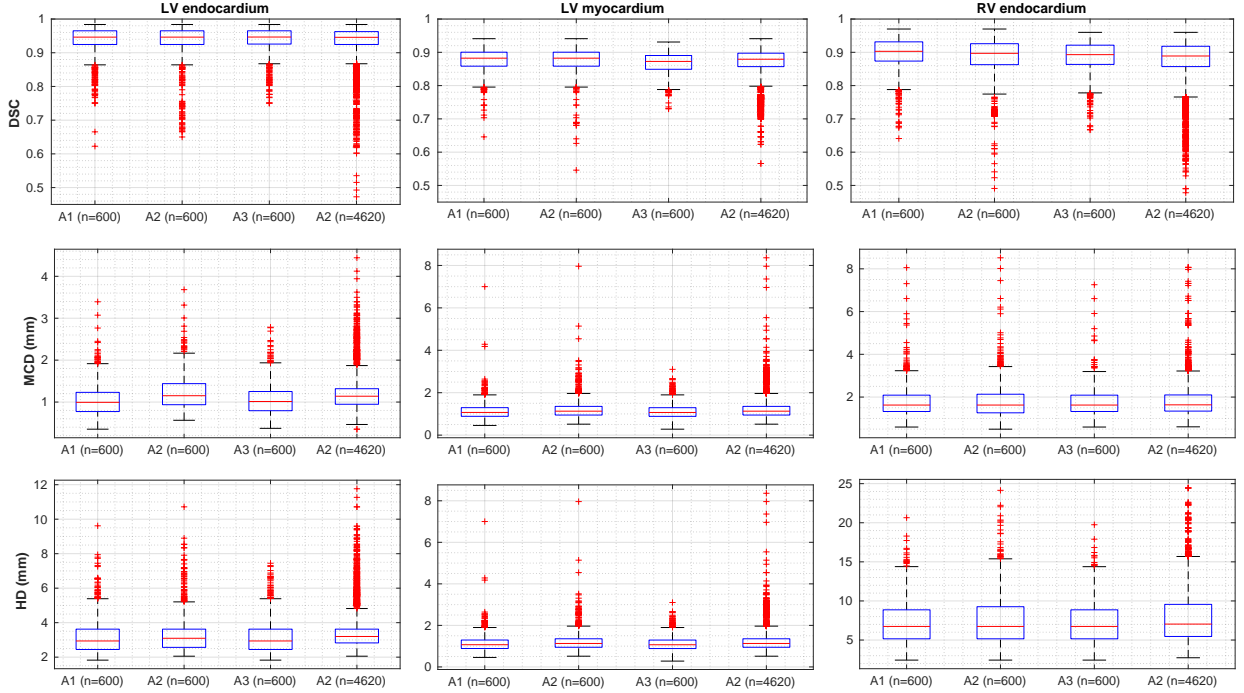


Figure 4: Segmentation accuracy expressed in terms of DSC, MCD and HD.

Table 5: Subjects used for the analysis in Section 3.2 after quality control.

Datasets	Total number of subjects (n)	Detected by IQA only	Detected by SQA only	Detected by IQA & SQA	Remain for analysis
Healthy population [7]	800	0	21	0	779
All manually segmented	4,620	145	105	60	4,430
Dataset used in [8]	600	0	11	0	589
UKB dataset	20,000	284	234	138	19,620

Figure 5 shows the Bland-Altman (top) and correlation (bottom) plots of the ventricular parameters computed based on the proposed automatic method and the manual reference on 4620 test subjects. The Bland-Altman plot is commonly used for analysing agreement and bias between two measurements. The Bland-Altman plots show good limits of agreement and also the mean difference line nearly at zero, which suggests that the clinical indexes obtained through the automatic approach have little bias. By contrast, the bias between different pairs of human observers as reported in [8] is considerable – nearly 8 (ml) for LVEDV and LVESV, about 8 (g) for LVM, and around 15 (ml) for RVEDV and RVESV.

Figure 5 shows the correlation plots between manual and automatic methods for the different

cardiac function indexes. The correlation coefficient (corr) is a measure of the strength of the relationship between two sets of observations. The strength and direction of the relationship indicates the predictive power of our framework. Coefficients for all indexes range from 0.85 to 0.91 indicating a strong relationship between the manual and automatic approaches.

To illustrate whether the values of clinical indexes computed automatically share the same distribution as the manual approach, we visualise their distributions. Figure 6 shows probability distribution plots (top) and Q-Q plots (bottom) for various cardiac functional indexes computed both manually and automatically over the full cohort where manual segmentations are available. On the plots, the distribution of the various indexes closely match

Table 6: The difference in clinical measures derived from the proposed method and manual segmentation. GT: ground truth provided by manual segmentation [7]. Values indicate mean \pm standard deviation.

	GT (n=800)	Automatic (n=800)	GT (n=4620)	Automatic (n=4620)	Automatic (n=20000)
LVEDV (ml)	144 \pm 34	146 \pm 31	144 \pm 34	144 \pm 33	142 \pm 26
LVESV (ml)	59 \pm 18	60 \pm 18	59 \pm 20	60 \pm 23	53 \pm 14
LVSV (ml)	85 \pm 20	86 \pm 18	84 \pm 18	84 \pm 19	89 \pm 18
LVEF (%)	60 \pm 6	60 \pm 7	60 \pm 6	59 \pm 7	63 \pm 6
LVM (g)	86 \pm 24	87 \pm 23	88 \pm 23	91 \pm 23	92 \pm 18
RVEDV (ml)	154 \pm 40	154 \pm 40	152 \pm 37	160 \pm 49	165 \pm 41
RVESV (ml)	69 \pm 24	71 \pm 26	67 \pm 22	77 \pm 26	61 \pm 24
RVSV (ml)	85 \pm 20	83 \pm 21	84 \pm 18	82 \pm 24	90 \pm 27
RVEF (%)	56 \pm 6	54 \pm 7	57 \pm 6	54 \pm 11	60 \pm 9

Table 7: The difference in clinical measures between the automatic and manual segmentations, as well between measurements by different human observers. M: ground truth provided by manual segmentation [7]. VF: Automatic segmentation obtained from the automatic segmentation software inlineVF D13A. Values indicate mean \pm standard deviation.

(a) Absolute difference										
	O1 vs O2 (n=50)	O2 vs O3 (n=50)	O3 vs O1 (n=50)	A1 vs M (n=50)	A2 vs M (n=50)	A1 vs M (n=600)	VF vs M (n=600)	A2 vs M (n=600)	A3 vs M (n=600)	A2 vs M (n=4620)
LVEDV	6.1 \pm 4.4	8.8 \pm 4.8	4.8 \pm 3.1	4.3 \pm 4.9	5.9 \pm 4.2	6.1 \pm 5.3	12.4 \pm 18.5	7.9 \pm 9.1	6.5 \pm 5.4	9.9 \pm 7.5
LVESV	4.1 \pm 4.2	6.7 \pm 4.2	7.1 \pm 3.8	6.5 \pm 5.4	6.8 \pm 5.1	5.3 \pm 4.9	9.2 \pm 14.8	7.0 \pm 10.0	5.1 \pm 5.0	8.2 \pm 6.3
LVM	4.2 \pm 3.2	6.6 \pm 4.9	6.5 \pm 4.8	6.4 \pm 3.5	6.0 \pm 4.4	6.9 \pm 5.5	NA	7.1 \pm 6.3	7.0 \pm 5.4	9.0 \pm 6.7
RVEDV	11.1 \pm 7.2	6.2 \pm 4.6	8.7 \pm 5.8	8.4 \pm 6.8	10.0 \pm 5.8	8.5 \pm 7.1	NA	10.1 \pm 7.2	8.4 \pm 7.8	12.9 \pm 9.8
RVESV	15.6 \pm 7.8	6.6 \pm 5.5	11.7 \pm 6.9	13.9 \pm 9.9	10.0 \pm 6.5	7.2 \pm 6.8	NA	8.7 \pm 9.5	7.7 \pm 6.5	12.2 \pm 9.6

(b) Relative difference (%)										
	O1 vs O2 (n=50)	O2 vs O3 (n=50)	O3 vs O1 (n=50)	A1 vs M (n=50)	A2 vs M (n=50)	A1 vs M (n=600)	VF vs M (n=600)	A2 vs M (n=600)	A3 vs M (n=600)	A2 vs M (n=4620)
LVEDV	4.2 \pm 3.1	6.3 \pm 3.3	3.4 \pm 2.2	2.9 \pm 3.6	4.2 \pm 3.0	4.1 \pm 3.5	8.8 \pm 12.9	5.0 \pm 3.3	4.7 \pm 3.3	7.0 \pm 5.2
LVESV	6.8 \pm 7.5	12.5 \pm 8.5	11.7 \pm 5.1	12.5 \pm 11.2	10.2 \pm 8.1	9.5 \pm 9.5	17.0 \pm 27.7	10.2 \pm 9.6	9.3 \pm 9.4	12.2 \pm 9.6
LVM	4.4 \pm 3.3	6.0 \pm 3.7	6.7 \pm 4.6	8.0 \pm 4.8	6.5 \pm 4.1	8.3 \pm 7.6	NA	8.1 \pm 8.2	8.3 \pm 7.7	8.2 \pm 7.6
RVEDV	8.0 \pm 5.0	4.2 \pm 3.1	5.7 \pm 3.6	5.7 \pm 4.3	7.3 \pm 4.2	5.6 \pm 4.6	NA	6.2 \pm 5.0	5.4 \pm 4.7	7.8 \pm 5.1
RVESV	30.6 \pm 15.5	10.9 \pm 8.3	16.9 \pm 9.2	29.8 \pm 22.1	22.0 \pm 8.4	11.8 \pm 12.2	NA	16.1 \pm 9.7	12.4 \pm 9.0	19.4 \pm 15.0

those obtained from the manual segmentations– a common distribution, common location and scale, similar distributional shapes, and similar tail behaviour.

We also computed the regional LV myocardial wall parameters in terms of thickness, thickening, and motion. Figure 7 shows the mean and standard deviation of the regional analysis of 4620 subjects in automatic and manual approaches in a bullseye display, based on the AHA 17-Segment model. It is clear that the two panels (top and bottom) are very similar in most regions in terms of the mean and standard deviation, which confirms the quality of the fully automatic pipeline. Indeed, results already published in many clinical journals [21, 22, 23, 24, 25, 26, 27, 28], which are mainly based on manual

delineation of a few dozen images confirm the values ranges obtained and presented in our bullseye plots.

Figures 8 and 9 show the distribution of wall thickness, thickening and motion for all AHA-17 segments in the LV myocardium. The histograms show measurements from automatic segmentation for two cohorts ($n = 4620$ and $n = 20K$) as well as from manual delineations. The figures show excellent agreement between measurements from automatic segmentations from both cohorts and those derived from manual delineations.

Furthermore, we performed two-sample Kolmogorov-Smirnov (K-S) tests to show that ventricular parameters obtained through manual and automatic approaches are drawn from the same population, under the null hypothesis that

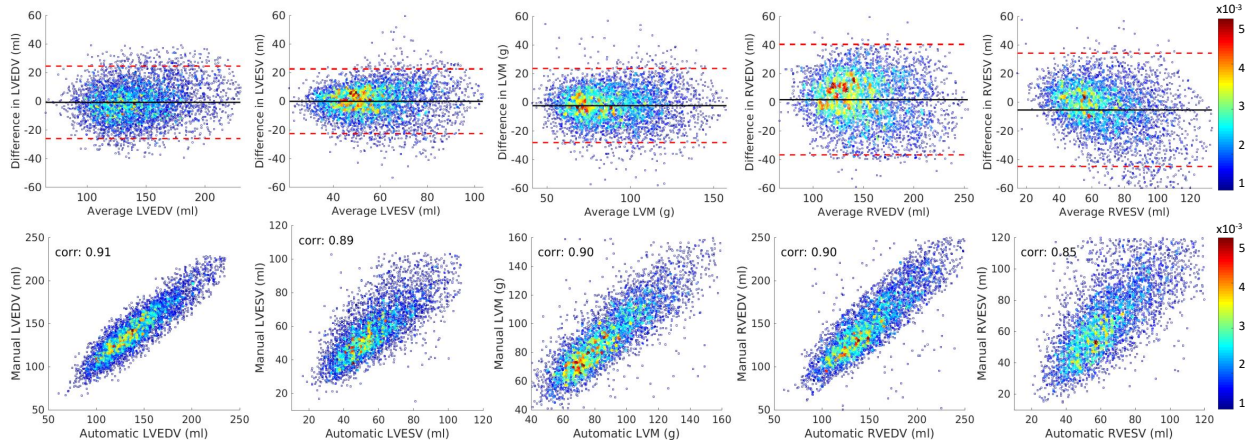


Figure 5: Repeatability of various cardiac functional indexes: manual vs automatic analysis on **4620** subjects from the UK Biobank cohort. First row shows the **Bland-Altman** plots for various cardiac functional indexes computed both manually and automatically where manual segmentation is available. The black line denotes the mean difference (bias) and the two red dashed lines denote ± 1.96 standard deviations from the mean. Second row shows the **Correlation** plots for various cardiac functional indexes computed both manually and automatically, where manual segmentation is available.

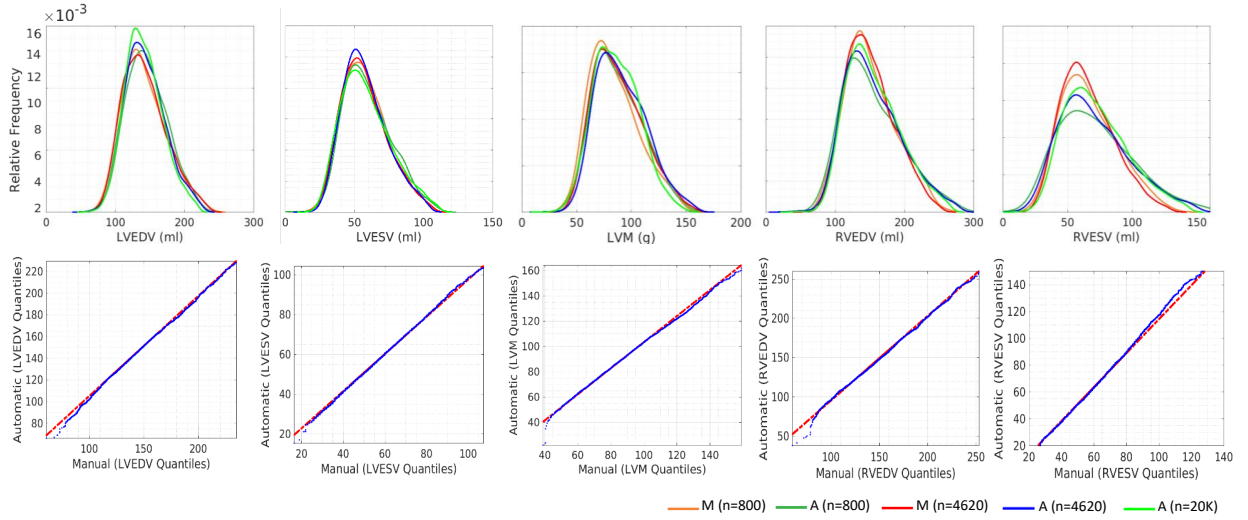


Figure 6: Distribution of various cardiac functional indexes: manual vs automatic analysis on 4620 subjects. First and second row show the **Probability Distribution** and **Q-Q** plots, respectively for various cardiac functional indexes computed both manually and automatically where manual segmentation is available.

the manual and automatic methods are from the same continuous distribution in terms of clinical indexes. K-S test on different global and regional indexes does not reject the null hypothesis of being from the same distribution at the 5% significance level.

Based on the exhaustive manual quantification provided by the authors in [7] and [8], where inter-observer variability is reported between three expert raters, it is evident that our automatic quan-

tification framework is well within human inter-observer variability ranges. On Table 7 it can be seen that our workflow’s errors for different indexes reported in [7] and [8] are well below maximum variability values.

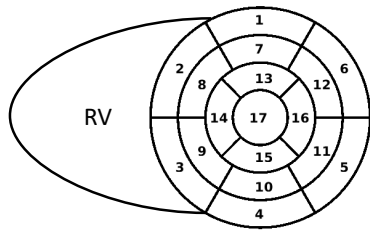
4. Conclusion

In this paper, we propose a fully automatic workflow capable of performing high throughput end-

to-end 3D cardiac image analysis. We tested our workflow on a reference cohort of 4620 subjects for which manual delineations, and reference functional indexes exist. Our results show statistically significant agreement between the manually obtained global/regional reference indexes (through extensive and laborious contouring of a large dataset), and those computed automatically using the proposed workflow.

5. ACKNOWLEDGMENTS

R. Attar was funded by the Faculty of Engineering Doctoral Academy Scholarship, University of Sheffield. This work has been partially supported by the MedIAN Network (EP/N026993/1) funded by the Engineering and Physical Sciences Research Council (EPSRC), and the European Commission through FP7 contract VPH-DARE@IT (FP7-ICT-2011-9-601055) and H2020 Program contract InSilc (H2020-SC1-2017-CNECT-2- 777119). The UKB CMR dataset has been provided under UK Biobank Application 2964.



AHA 17-Segment Model

AHA 17-segments:

1. basal anterior
2. basal anteroseptal
3. basal inferoseptal
4. basal inferior
5. basal inferolateral
6. basal anterolateral
7. mid anterior
8. mid anteroseptal

9. mid inferoseptal
10. mid inferior
11. mid inferolateral
12. mid anterolateral
13. apical anterior
14. apical septal
15. apical inferior
16. apical lateral
17. apex

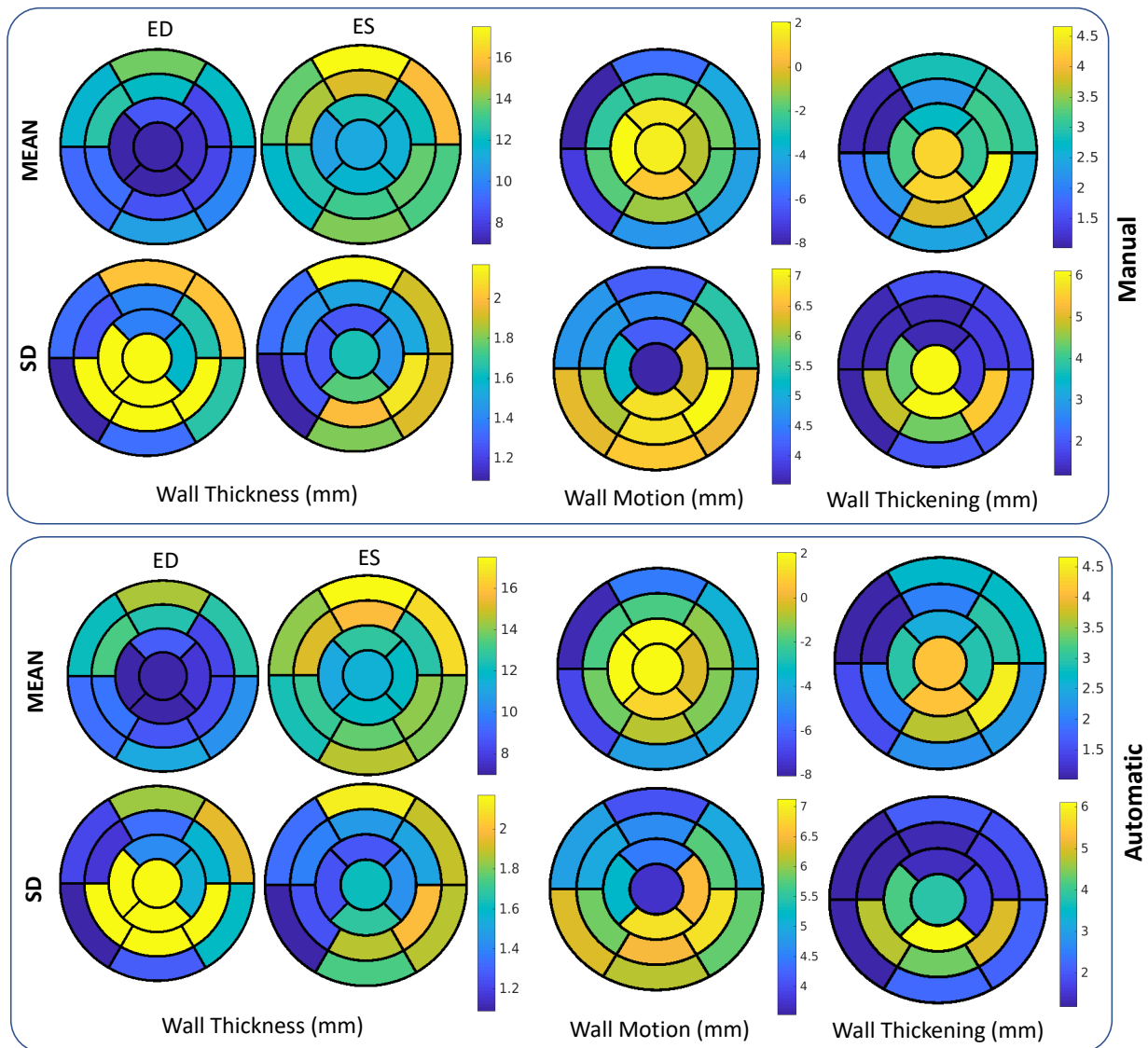


Figure 7: Segmental left ventricular parameters of 4620 subjects presented in bulls-eye display.

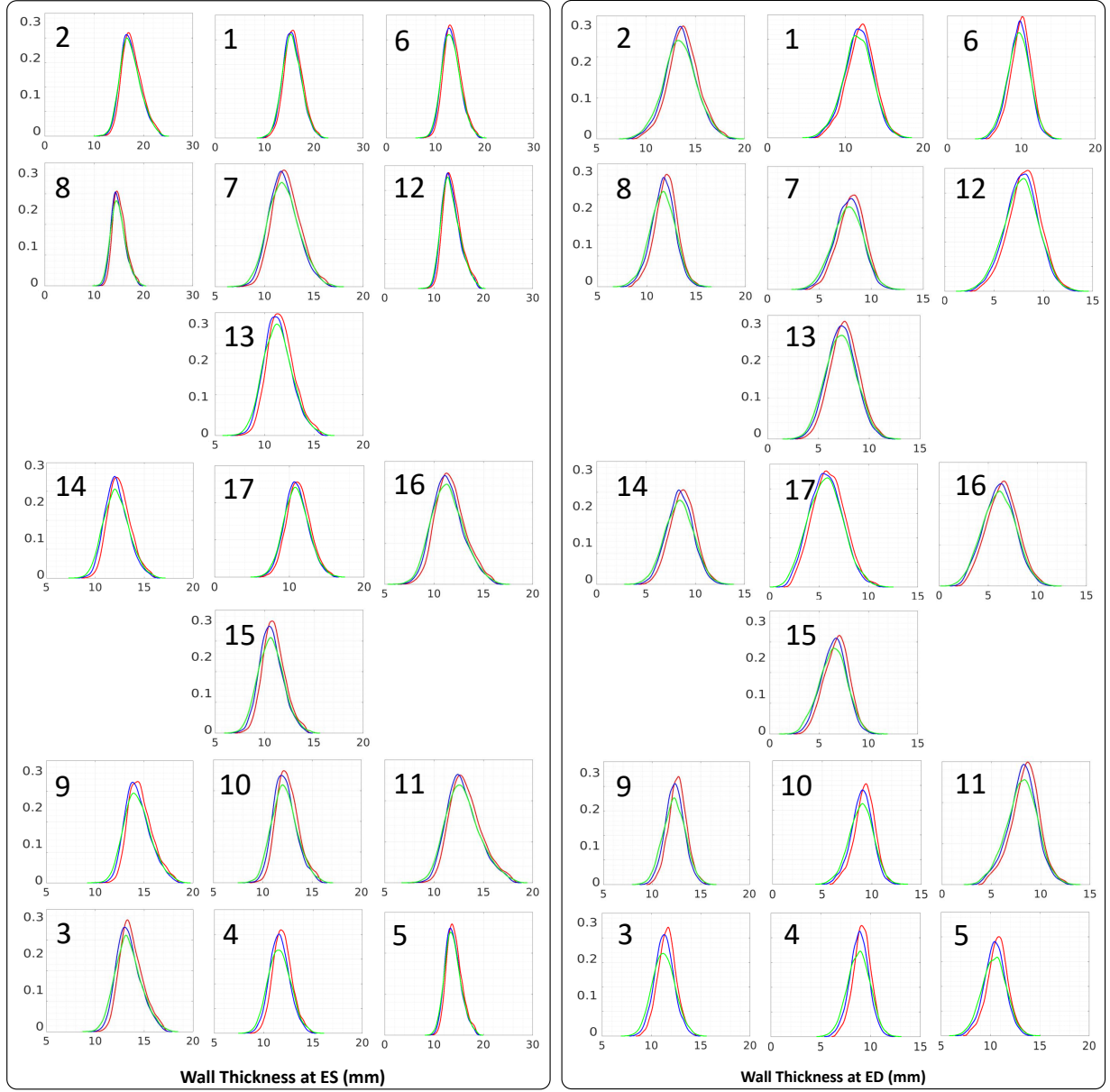


Figure 8: Regional analysis of LV shapes on 20K subjects in terms of distribution of wall thickness at ED and ES phases, wall motion and thickening. Red, blue, green lines indicate ground truth values (4620 subjects), automatic values (4620 subjects) and automatic values (20K subjects), respectively.

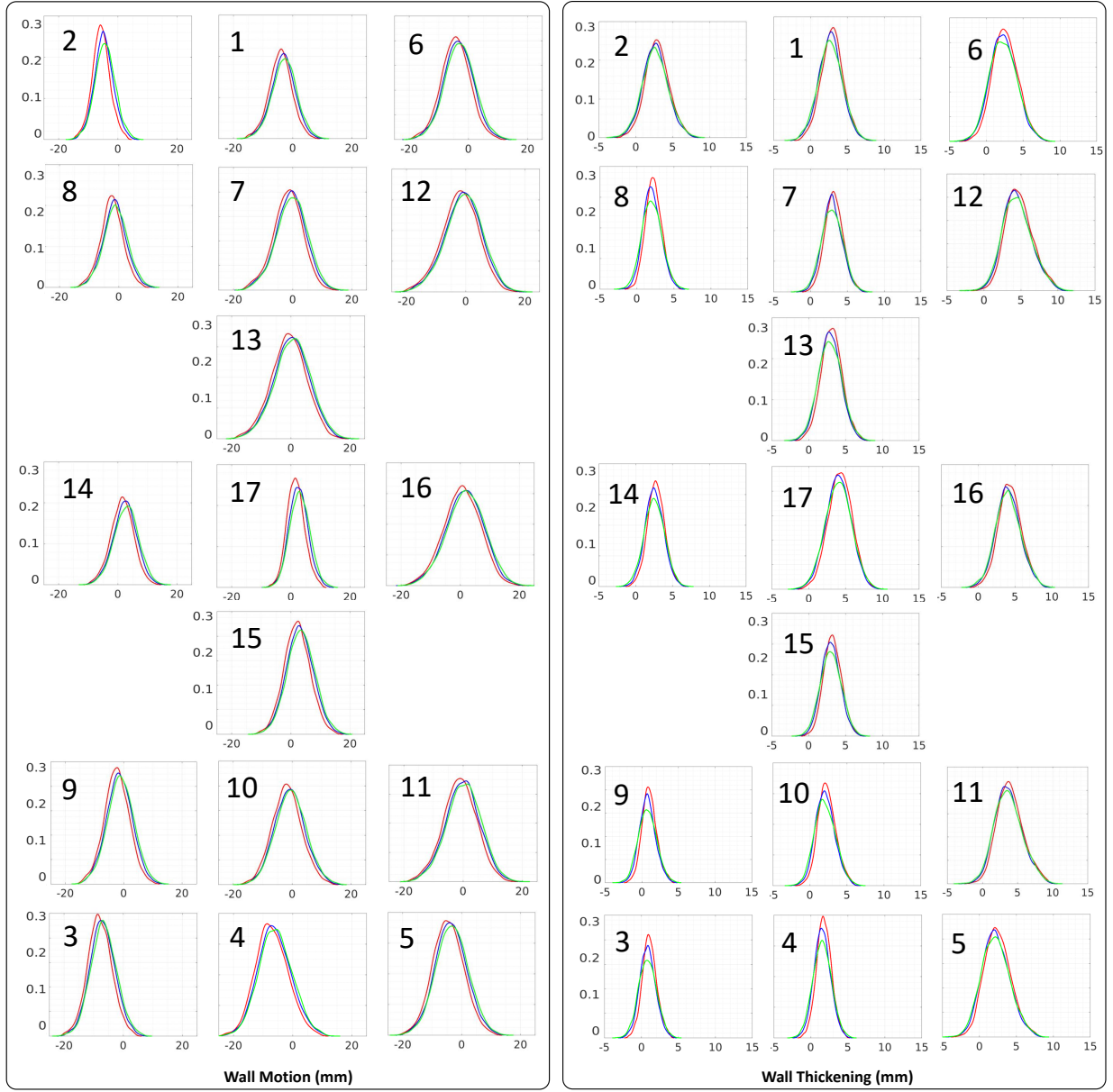


Figure 9: Regional analysis of LV shapes on 20K subjects in terms of distribution of wall thickness at ED and ES phases, wall motion and thickening. Red, blue, green lines indicate ground truth values (4620 subjects), automatic values (4620 subjects) and automatic values (20K subjects), respectively.

References

- [1] G. A. Roth, C. Johnson, A. Abajobir, F. Abd-Allah, S. F. Abera, G. Abyu, M. Ahmed, B. Aksut, T. Alam, K. Alam, et al., Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015, *Journal of the American College of Cardiology* 70 (2017) 1–25.
- [2] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, S. Iyengar, Computational health informatics in the big data age: a survey, *ACM Computing Surveys (CSUR)* 49 (2016) 12.
- [3] P. Medrano-Gracia, B. R. Cowan, A. Suinesiaputra, A. A. Young, Challenges of cardiac image analysis in large-scale population-based studies, *Current cardiology reports* 17 (2015) 9.
- [4] A. Lardo, Z. A. Fayad, N. Chronos, V. Fuster, *Cardiovascular magnetic resonance: established and emerging applications*, Taylor & Francis, 2004.
- [5] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt, et al., UK Biobanks cardiovascular magnetic resonance protocol, *Journal of cardiovascular magnetic resonance* 18 (2015) 8.
- [6] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, et al., Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK biobank-rationale, challenges and approaches, *Journal of Cardiovascular Magnetic Resonance* 15 (2013) 46.
- [7] S. E. Petersen, N. Aung, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, J. M. Francis, M. Y. Khanji, E. Lukaschuk, A. M. Lee, et al., Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in caucasians from the UK Biobank population cohort, *Journal of Cardiovascular Magnetic Resonance* 19 (2017) 18.
- [8] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, et al., Automated cardiovascular magnetic resonance image analysis with fully convolutional networks., *Journal of Cardiovascular Magnetic Resonance* (2018).
- [9] K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, S. S. Ghosh, Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python, *Frontiers in neuroinformatics* 5 (2011) 13.
- [10] V. Klinker, S. Muzzarelli, N. Lauriers, D. Locca, G. Vincenti, P. Monney, C. Lu, D. Nothnagel, G. Pilz, M. Lombardi, et al., Quality assessment of cardiovascular magnetic resonance in the setting of the european CMR registry: description and validation of standardized criteria, *Journal of Cardiovascular Magnetic Resonance* 15 (2013) 55.
- [11] L. Zhang, A. Gooya, B. Dong, R. Hua, S. E. Petersen, P. Medrano-Gracia, A. F. Frangi, Automated quality assessment of cardiac MR images using convolutional neural networks, in: *International Workshop on Simulation and Synthesis in Medical Imaging*, Springer, pp. 138–145.
- [12] V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, B. Glocker, Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth, *IEEE Transactions on Medical Imaging* (2017).
- [13] X. Albà, K. Lekadir, M. Pereañez, P. Medrano-Gracia, A. A. Young, A. F. Frangi, Automatic initialization and quality control of large-scale cardiac mri segmentations, *Medical image analysis* 43 (2018) 129–141.
- [14] H. C. Van Assen, M. G. Danilouchkine, A. F. Frangi, S. Ordás, J. J. Westenberg, J. H. Reiber, B. P. Lelieveldt, SPASM: a 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data, *Medical Image Analysis* 10 (2006) 286–303.
- [15] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, Active shape models-their training and application, *Computer vision and image understanding* 61 (1995) 38–59.
- [16] G. V. Heller, M. D. Cerqueira, N. J. Weissman, V. Dilsizian, A. K. Jacobs, S. Kaul, W. K. Laskey, D. J. Pennell, J. A. Rumberger, T. Ryan, et al., Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association, *Journal of Nuclear Cardiology* 9 (2002) 240–245.
- [17] A. F. Frangi, W. J. Niessen, M. A. Viergever, Three-dimensional modeling for functional analysis of cardiac images, a review, *IEEE transactions on medical imaging* 20 (2001) 2–5.
- [18] F. H. Sheehan, E. L. Bolson, H. T. Dodge, D. G. Mathey, J. Schofer, H. Woo, Advantages and applications of the centerline method for characterizing regional ventricular function., *Circulation* 74 (1986) 293–305.
- [19] D. E. Bild, D. A. Bluemke, G. L. Burke, R. Detrano, A. V. Diez Roux, A. R. Folsom, P. Greenland, D. R. Jacobs Jr, R. Kronmal, K. Liu, et al., Multi-ethnic study of atherosclerosis: objectives and design, *American journal of epidemiology* 156 (2002) 871–881.
- [20] Q. Zheng, H. Delingette, N. Duchateau, N. Ayache, 3D consistent & robust segmentation of cardiac images by deep learning with spatial propagation, *IEEE Transactions on Medical Imaging* (2018).
- [21] F. Andre, S. Lehrke, H. A. Katus, H. Steen, Reference values for the left ventricular wall thickness in cardiac MRI in a modified AHA 17-segment model, *Journal of Cardiovascular Magnetic Resonance* 14 (2012) P223.
- [22] A. Deviggiano, P. Carrascosa, M. De Zan, C. Capuñay, H. Deschle, G. A. Rodríguez Granillo, Wall thickness and patterns of fibrosis in hypertrophic cardiomyopathy assessed by cardiac magnetic resonance imaging, *Revista Argentina de Cardiología* 84 (2016).
- [23] V. O. Puntmann, Y. G. Yap, W. McKenna, A. J. Camm, Significance of maximal and regional left ventricular wall thickness in association with arrhythmic events in patients with hypertrophic cardiomyopathy, *Circulation Journal* 74 (2010) 531–537.
- [24] R. E. Kanza, H. Higashino, T. Kido, A. Kurata, M. Saito, Y. Sugawara, T. Mochizuki, Quantitative assessment of regional left ventricular wall thickness and thickening using 16 multidetector-row computed tomography: comparison with cine magnetic resonance imaging, *Radiation medicine* 25 (2007) 119–126.
- [25] M. Prasad, A. Ramesh, P. Kavanagh, B. K. Tamarappoo, R. Nakazato, J. Gerlach, V. Cheng, L. E. Thomson, D. S. Berman, G. Germano, et al., Quantification

- of 3D regional myocardial wall thickening from gated magnetic resonance images, *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 31 (2010) 317–327.
- [26] F. Le Ven, K. Bibeau, É. De Larochellière, H. Tizón-Marcos, S. Deneault-Bissonnette, P. Pibarot, C. F. Deschepper, É. Larose, Cardiac morphology and function reference values derived from a large subset of healthy young Caucasian adults by magnetic resonance imaging, *European Heart Journal-Cardiovascular Imaging* 17 (2015) 981–990.
- [27] A. Baltabaeva, M. Marciniak, B. Bijmens, J. Moggridge, F. J. He, T. F. Antonios, G. A. MacGregor, G. R. Sutherland, Regional left ventricular deformation and geometry analysis provides insights in myocardial remodelling in mild to moderate hypertension, *European Journal of Echocardiography* 9 (2007) 501–508.
- [28] I. Codreanu, T. J. Pegg, J. B. Selvanayagam, M. D. Robson, O. J. Rider, C. A. Dasanu, B. A. Jung, D. P. Taggart, S. J. Golding, K. Clarke, et al., Normal values of regional and global myocardial wall motion in young and elderly individuals using navigator gated tissue phase mapping, *Age* 36 (2014) 231–241.