Quantitative CMR Population Imaging on 20,000 Subjects of the UK Biobank Imaging Study: LV/RV Quantification Pipeline and its Evaluation

Rahman Attar[a,b,*], Marco Pereañez[a,b], Ali Gooya[a], Xènia Albà[d], Le Zhang[c], Milton Hoz de Vila[a], Aaron M. Lee[f,g], Nay Aung[f,g], Elena Lukaschuk[e], Mihir M. Sanghvi[f,g], Kenneth Fung[f,g], Jose Miguel Paiva[f,g], Stefan K. Piechnik[e], Stefan Neubauer[e], Steffen E. Petersen[f,g], Alejandro F. Frangi[a,b,*]

[a] *Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK.*
[b] *Biomedical Imaging Department, Leeds Institute for Cardiovascular and Metabolic Medicine (LICAMM), School of Medicine, University of Leeds, Leeds, UK.*
[c] *Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, UK.*
[d] *Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Universitat Pompeu Fabra, Barcelona, Spain.*
[e] *Oxford Centre for Clinical Magnetic Resonance Research (OCMR), Division of Cardiovascular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK.*
[f] *William Harvey Research Institute, NIHR Barts Biomedical Research Unit, Queen Mary University of London, London, UK*
[g] *Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, UK*

Abstract

Population imaging studies generate data for developing and implementing personalised health strategies to prevent, or more effectively treat disease. Large prospective epidemiological studies acquire imaging for pre-symptomatic populations. These studies enable the early discovery of alterations due to impending disease, and enable early identification of individuals at risk. Such studies pose new challenges requiring automatic image analysis. To date, few large-scale population-level cardiac imaging studies have been conducted. One such study stands out for its sheer size, careful implementation, and availability of top quality expert annotation; the UK Biobank (UKB). The resulting massive imaging datasets (targeting ca. 100,000 subjects) has put published approaches for cardiac image quantification to the test. In this paper, we present and evaluate a cardiac magnetic resonance (CMR) image analysis pipeline that properly scales up and can provide a fully automatic analysis of the UKB CMR study. Without manual user interactions, our pipeline performs end-to-end image analytics from multi-view cine CMR images all the way to anatomical and functional bi-ventricular quantification. All this, while maintaining relevant quality controls of the CMR input images, and resulting image segmentations. To the best of our knowledge, this is the first published attempt to fully automate the extraction of global and regional reference ranges of all key functional cardiovascular indexes, from both left and right cardiac ventricles, for a population of 20,000 subjects imaged at 50 time frames per subject, for a total of one million CMR volumes. In addition, our pipeline provides 3D anatomical bi-ventricular models of the heart. These models enable the extraction of detailed information of the morphodynamics of the two ventricles for subsequent association to genetic, omics, lifestyle habits, exposure information, and other information provided in population imaging studies. We validated our proposed CMR analytics pipeline against manual expert readings on a reference cohort of 4,620 subjects with contour delineations and corresponding clinical indexes. Our results show broad significant agreement between the manually obtained reference indexes, and those automatically computed via our framework. 80.67% of subjects were processed with mean contour distance of less than 1 pixel, and 17.50% with mean contour distance between 1 and 2 pixels. Finally, we compare our pipeline with a recently published approach reporting on UKB data, and based on deep learning. Our comparison shows similar performance in terms of segmentation accuracy with respect to human experts.

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

1. Introduction

Cardiovascular disease (CVD) is the most prevalent cause of death worldwide (Roth et al., 2017). Diagnosis of CVDs is often made at late symptomatic stages, leading to late interventions at high cost and with substantially decreased efficacy of treatment. Early quantitative assessment of cardiac function that allows for proper preventive care, and early cardiovascular treatment is therefore paramount. To support such an approach, large-scale population-based imaging studies of CVDs are increasingly possible given the advent of standardised robust non-invasive imaging methods, and the infrastructure for big data analysis (Fang et al., 2016). These advancements open further opportunities for gaining new information about the development and progression of CVDs across various population groups (Lardo et al., 2004; Medrano-Gracia et al., 2015).

The analysis and interpretation of cardiac structural and functional indexes in large-scale population imaging data can help identify patterns and trends across population groups, and accordingly, reveal insights into key risk factors before CVDs fully develop. Established to investigate the determinants of a disease, the UK Biobank (UKB) is one of the world's largest prospective population studies (Petersen et al., 2015). The UKB data contain extensive baseline questionnaire data, biological samples, physical measurements, and cardiovascular magnetic resonance (CMR) images to establish cardiovascular imaging-derived phenotypes (Petersen et al., 2013). CMR is an important component of multi-organ multi-modality imaging visits for patients in multiple dedicated UKB imaging centres that will acquire 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 by Petersen et al. (2017), in which CMR scans were manually delineated and analysed by a team of eight expert observers using the commercially available cvi42 post-processing software (Version 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Canada). The expert team comprised of biomedical engineers, radiologists, image analysts and cardiologists, evaluated the quality of every image, and performed delineations. In cases where the image quality was doubtful, the team jointly decided upon exclusion. These reference values (delineations and volumes) comprise 4,620 subjects and are used in our present study to validate our proposed framework and workflow.

In this paper, we present a novel fully automatic 3D image parsing workflow with embedded quality control, and evaluate its performance on the UKB. We validate our results by comparing with published manual analysis and one state-of-the-art method. Our proposed workflow is capable of segmenting the cardiac ventricles and generating global and regional clinical reference ranges comparable to those obtained by human raters and flagship methods.

In addition to comparing against manual measurements, we also compare our performance against one state-of-the-art method, i.e., the recent work by Bai et al. (2018) in which the authors propose a 2D convolutional neural network (CNN)-based segmentation method for analysis of the UKB CMR images. Though in our study, we processed a much greater number of subjects $(20,000)$, we performed experiments with smaller subsets of data to make direct comparisons with the existing literature. We are interested in showing the advantages of true 3D shape analysis, over 2D CNN-based techniques, which, due to their per-slice disjoint nature, and absence of global constraints, lack the ability to infer or extrapolate noisy or missing data. We believe true 3D analysis is valuable, or even essential, for further structural analysis of regional myocardial function. Our 3D generative-based approach ensures global coherence of the cardiac anatomy and naturally lends itself to further analysis in which full 3D anatomy is necessary; for example, in mechanical and flow simulations.

Finally, since the power of population studies lies in the ability to provide normative reference values for sub-populations, enabling more patient-specific evaluation, we provide reference ranges for cardiac clinical indexes in sub-populations based on age-group and gender.

The main contributions of this paper are, first, reproducing the cardiac functional index ranges derived from expert delineations reported in (Petersen et al., 2015), and providing additional 3D-based ranges of local variation. Second, showcasing a

[*]Corresponding author
Email addresses: r.attar@leeds.ac.uk (Rahman Attar), a.frangi@leeds.ac.uk (Alejandro F. Frangi)

fully scalable framework, capable of processing arbitrarily large population imaging studies, in a completely automatic manner. In this paper we demonstrate this by processing 20,000 subjects from the UKB study, each comprised of 50 time frames for a total of one million image volumes, starting from raw input data, through data cleaning, quality assessment, 3D segmentation, volume computation, and statistical analysis.

The remainder of this paper is organised as follows. In Section 2, we present our strategy for data processing scalability, and detail each of the modules comprising our image quantification pipeline. In Section 3, we present a thorough evaluation of our pipeline, both from technical, and clinical perspectives, including detailed statistics on global and local cardiovascular indexes. Finally, in Section 4, we present final remarks.

2. Methodology

Illustrated in Figure 1, our CMR image parsing pipeline consists of the following four phases: (1) pre-processing; (2) quality analysis; (3) segmentation; and (4) quantification. In the subsections that follow, we describe the methods used within each step and our design choices. In the next subsection, we highlight the framework used to integrate this pipeline and streamline its execution both in terms of scalability and distributability.

2.1. Workflow Integration and Execution

To scale both data access and computation, we propose a modular pipeline and developed an inhouse cloud-based image analytics framework called MULTI-X [1] (de Vila et al., 2018). MULTI-X enables both distributed access to data storage and distributed execution of image analysis pipelines on the cloud. Further, MULTI-X facilitates secure access and execution, component integration and interoperability (e.g., across different programming languages, frameworks, operating systems, and hardware), workflow execution, monitoring, and execution report generation. MULTI-X can also serve as middleware between storage and computing cloud providers (e.g., Amazon Web Services, GoogleCloud, and Microsoft Azure), workflow managers (e.g., Taverna and Nipype), data sources (e.g., UKB servers) and analytics tools providers. In our implementation, we selected Nipype (Gorgolewski et al., 2011) as the workflow manager. Further, we selected Amazon Web Services[2] to provide high-performance storage and computing in a cloud-based environment. More specifically, an Amazon Simple Storage Service (S3) provided unstructured data storage, Amazon Redshift provided data warehousing for petabyte-scale data analysis, and Amazon's Elastic Cloud Computing (EC2) enabled on-demand adaptive cloud computing.

2.2. Data Pre-processing

Before describing our data pre-processing phase, first note that we accessed the UKB data under Access Applications #2964 and #11350. UKB project has been approved by National Research Ethics Service North West (11/NW/0382) and all participants gave written consent to participate and to publish as part of the UKB recruitment process. The full CMR protocol in the UKB has been described in detail elsewhere (Petersen et al., 2015). Researchers can apply to use the UKB resource for health related research that is in the public interest[3].

Once obtained, data were transferred to a secure AWS S3 server accessible from an experimental deployment of MULTI-X, the aforementioned cloud-based infrastructure for our pipeline-oriented image analytics. A production deployment of MULTI-X installed within the UKB is presently under way and will provide a corporate data analytics framework. When new data are available, this production implementation of MULTI-X will provide continued image analytics for the volunteers whose CMR data collection is an ongoing task. The UKB Imaging Study undertakes detailed MRI scans of key vital organs of the human body using specialised imaging protocols that extend CMR. For each volunteer, relevant CMR subseries are extracted from the full imaging study, viz. short axis (SAX) and long axis (LAX) two-, three- and four-chamber CMR images.

2.3. Quality Analysis

At least two quality analysis modules are required to ensure the reliability of the extracted cardiac indexes. The first module assesses the quality of the input images, whereas the second module assesses the quality of the quantification outputs, i.e., the

[1] https://multi-x.org

[2] https://aws.amazon.com

[3] https://www.ukbiobank.ac.uk/register-apply

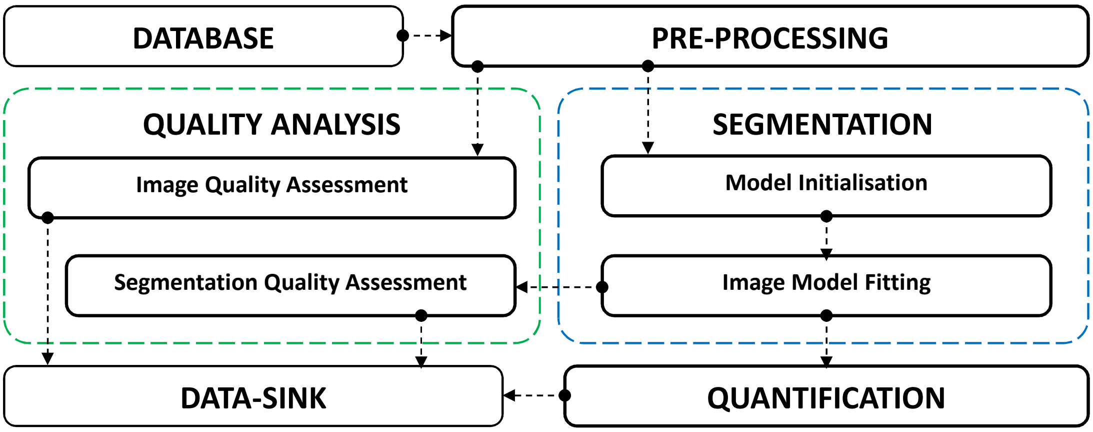


Figure 1: Schematic showing our fully automatic image parsing framework for large-scale analysis of cardiac ventricles. CMR images first go through the pre-processing phase, then flow into both the quality analysis and segmentation phases, which in turn communicate with one another, finally producing output that the last phase of quantification handles.

generated 3D segmentations. Each of these is described in the subsections that follow.

2.3.1. Image Quality Assessment

Despite careful and strict imaging protocols, a significant portion of the data collected in population imaging studies, inevitably falls outside standard operating procedures. To ensure the quality and correctness of the collected data, thereby optimising the accuracy of the generated segmentation results, an image quality assessment (IQA) module detects suboptimal images whose inclusion in subsequent analysis would impair aggregated statistics over the entire cohort.

More specifically, because the absence of basal and/or apical slices in SAX views forms the most frequently occurring problem affecting the accuracy of volumetric measurements and corresponding clinical indexes (Klinke et al., 2013), our IQA module detects situations in which these slices are missing. In our design, SAX slices are processed independently through two CNN classifiers that determine the presence/absence of basal and apical slices, respectively. Details of the algorithm we used to achieve this effect can be found in the work published by Zhang et al. (2016).

2.3.2. Segmentation Quality Assessment

Regarding segmentation quality assessment (SQA), large anatomical variations found across subject populations (Valindria et al., 2017) and other forms of poor image quality beyond full ventricular coverage can cause image segmentation failures. We therefore propose an automated self-diagnosis mechanism for detecting unsatisfactory segmentation results. Flagged images can then be either re-processed with revised parameters or discarded from subsequent statistical analyses. We incorporate a segmentation quality assessment approach presented by Albà et al. (2018). The SQA module uses a random forest classifier trained to distinguish between successful and unsuccessful segmentations based on intensity features around the blood pool and myocardial boundaries.

2.4. Segmentation

For the segmentation phase of our workflow, we use SAX and LAX CMR images to estimate the approximate position and orientation of the cardiac ventricles. We then initialise the segmentation of the cardiac structure following a Sparse Active Shape Model (SPASM) approach (Van Assen et al., 2006). More specifically, 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 the frames showing maximum and minimum volumes, respectively. Before running our segmentation approach across all subjects, we applied grid search optimisation to a subset of 50 subjects to identify the parameters having the greatest impact on segmentation performance; we describe this further in Section 2.4.3.

2.4.1. Model Initialisation

To automatically initialise the model, we use the method proposed by Albà et al. (2018) with a further step to improve bi-ventricular model initialisation. First, the location of the LV is determined via a rough estimate of the intersection of slices from the SAX and LAX views. Next, a random forest regressor trained with two complementary feature descriptors (i.e., the Histogram of Oriented Gradients and Gabor Filters) predicts the landmark positions for the LV. We extend this to take into account image features corresponding to the RV, thereby improving the initial estimate for the location of the bi-ventricular heart. We then use these landmarks to estimate pose parameters that place a mean shape model near the heart. Finally, we use these pose parameters to initialise the first image volume in the set of images for the cardiac cycle (i.e., 50 cardiac phases). Subsequent time frames are automatically initialised via the shape model fitted to the immediately preceding cardiac phase.

2.4.2. Image Model Fitting

In this subsection, we consider how we fit the image model. First, the cardiac LV and RV segmentations are obtained via the aforementioned SPASM segmentation method that improves on the Active Shape Models (ASM) approach (Cootes et al., 1995) by addressing the sparsity found in imaging modalities such as CMR in which image information is sparsely distributed across the entirety of the image. The main components of the SPASM method are the Point Distribution Model (PDM), the Intensity Appearance Model (IAM) and a model matching algorithm.

The PDM encodes the mean and variance of the endocardial 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 generalised Procrustes-aligned shapes that preserve a 98% variance.

To illustrate this, assume a training set of M shapes, each described by N points in $\mathcal{R}^3$, i.e., $\mathbf{x}^i_j = (\mathbf{x}^i_j, \mathbf{y}^i_j, \mathbf{z}^i_j)$ with $i = 1, ..., M$ and $j = 1, ..., N$.

Further, let $\mathbf{s}_i = (\mathbf{x}^i_1, \mathbf{y}^i_1, \mathbf{z}^i_1, ..., \mathbf{x}^i_N, \mathbf{y}^i_N, \mathbf{z}^i_N)^T$ 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 $\mathbf{S} = [\mathbf{s}^1, ..., \mathbf{s}^M]$ be the set of all training shapes in matrix form. Here, all nuisance pose parameters (e.g., translation, rotation and scaling) have been removed from $\mathbf{S}$ using generalised Procrustes analysis. The shape class mean and covariance of $\mathbf{S}$ is then as follows:

$$\bar{\mathbf{s}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{s}_i \tag{1}$$

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

The shape covariance is represented in a low-dimensional space or PCA of the shape. That produces l eigenvectors $\mathbf{\Phi} = [\varphi_1\varphi_2...\varphi_l]$, and corresponding eigenvalues $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, ..., \lambda_l)$ of the covariance matrix computed via Singular Value Decomposition. 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:

$$\mathbf{s} \approx \bar{\mathbf{s}} + \mathbf{\Phi}\mathbf{b} \tag{3}$$

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

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

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

Next, for each landmark in $\mathbf{s}$, we build an IAM based on intensity information across all corresponding landmarks in all training shapes $\mathbf{s}_i$. More specifically, IAMs capture the local intensity distribution along cardiac boundaries. We proceed by sampling one-dimensional intensity profiles normal to the myocardial boundaries. Each profile has a length of $m = 15$ pixels. For the i-th landmark, we estimate mean intensity profile $\bar{\mathbf{g}}_i$ and corresponding image intensity covariance $\mathbf{S}_{g_i}$.

During image segmentation, the intersections of the current shape model instance with all image planes collectively define a stack of two-dimensional contours in $\mathcal{R}^3$. The algorithm proceeds by searching for the 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 minimise the Mahalanobis distance between a profile sampled at candidate position $\mathbf{y}_i$, $\mathbf{g}_i(\mathbf{y}_i)$ and corresponding model $\{\bar{\mathbf{g}}_i, \mathbf{S}_{g_i}\}$ as follows:

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

Given the sparse nature of CMR images, it is not uncommon during fitting to have mesh triangles that do not intersect with any image slices in the stack. In this situation, the points that comprise these triangles would not be updated or displaced by the IAM, instead, these points would be passively updated by fitting of the PDM. A mechanism that propagates displacements from points that are image-driven to nearby points that are not, is therefore necessary. SPASM implements a displacement propagation strategy modelled as a Gaussian kernel centred at any given image-driven point q by propagating its effect to a neighbouring point p based on Gaussian kernel

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

where σ is the width of the kernel. Having a non-zero Gaussian kernel is not an indispensable feature of the algorithm as non-image driven points would be indirectly updated by the PDM, nevertheless, this feature adds smoothness to the evolution of the surface mesh, and speeds up convergence of the algorithm.

2.4.3. Parameter Optimisation

SPASM segmentation is affected by four key parameters. We ran an exhaustive grid optimisation scheme to determine the best combination of parameters. The individual parameters and corresponding ranges that we tested were as follows:

1. Freedom of the PDM measured in standard deviations from the mean i.e. $\beta = 2, 2.5, 3$.
2. Length of the image sampling profile $\mathbf{g}_i$ used during image feature search, measured in pixels i.e. $l = 5, 10, 15$ pixels.
3. Standard deviation of the Gaussian kernel for the point displacement propagation feature i.e. $\sigma = 5, 7, 9$ mm.
4. Image orientations to use, i.e. using only SAX images or using both SAX and LAX during segmentation i.e. $v = SAX, ALL$.

Table 1 shows each of the 54 (i.e. $3 \times 3 \times 3 \times 2 = 54$) unique parameter combinations we used with our algorithm to segment 50 randomly selected subjects that had already been manually delineated by clinicians. Next, we computed the segmentation accuracy using three key metrics: Dice Similarity Coefficient (DSC), Mean Contour Distance (MCD) and Hausdorff Distance (HD). These metrics are defined on Equations 7, 8 and 9, respectively, in Section 3.1.

Table 1: The list of 54 distinct sets of segmentation parameters used in our segmentation algorithm parameter optimisation. As noted in the text, test 4 was the best choice.

Test	β	l	σ	v	**Test**	β	l	σ	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

Figure 2 shows three boxplots summarising the results based on the three metrics, for each of the 54 test parameter sets. The x-axis on each of the boxplots (DSC, MCD, HD) shows the test number, and the tests are sorted from best to worst performance. On Figure 2 it can been seen that the best performing parameter test set is test number 4, which appears at the left-most end of each of the plots. Specifically, the best parameter values were: $\beta = 2$, $l = 5$, $\sigma = 7$ and $v = ALL$. We used this parameter set for segmentation thereon.

2.5. Quantification

For the final phase of our workflow, we computed a thorough set of functional parameters based on blood-pool and myocardial volumes. To reproduce the reference ranges reported by Petersen

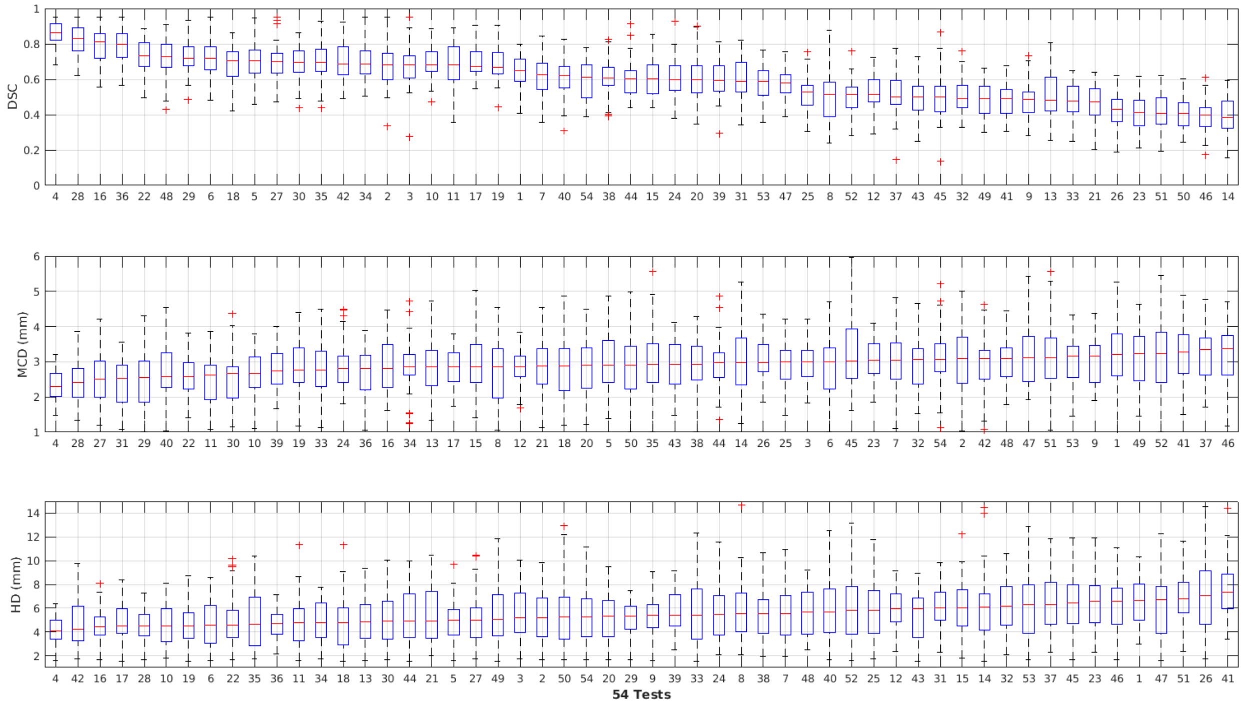


Figure 2: Results of our segmentation algorithm parameter optimisation. Here, the set of parameters that jointly yielded the best results for the DSC, MCD and HD metrics was test 4.

et al. (2017), our quantification module performs volume computations using Simpson's method of integration, whereby a cardiac 3D volume can be approximated by summing the areas within 2D segmentation contours, and multiplying by the interslice spacing. Because the output of our segmentation are 3D triangular meshes, before using Simpson's rule, we had to extract contours corresponding to the intersection between our segmentation and CMR image slices. The 3D model we use for segmentation is comprised of two structures; the LV and the RV. The LV is a closed water-tight mesh comprising both endocardial and epicardial walls. The RV is an open mesh representing only the RV endocardium. The RV has two openings, the atrio-ventricular valve opening, and pulmonary valve opening. The LV and RV sit adjacent to each other but are not connected.

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, wall motion and thickening computed for every segment in the AHA-17 cardiac subdivision scheme (Heller et al., 2002).

The global assessment of cardiac function is based on the following volumetric measurements (Frangi et al., 2001):

- End-Diastolic Volume (EDV) (ml): the volume of blood in the LV or RV before contraction. This is the highest ventricular volume of blood in the cardiac cycle.

- End-Systolic Volume (ESV) (ml): the volume of blood in the LV or RV at the end of contraction. This is the lowest ventricular volume of blood in the cardiac cycle.

- Stroke Volume (SV) (ml): the volume of blood pumped from the ventricle per beat obtained by subtracting the ESV from the EDV for a given ventricle. This term can be applied to either of the two ventricles.

- Ejection Fraction (EF) (%): the fraction of blood ejected from a ventricle of the heart with each heartbeat. This measure shows the pumping efficiency of the heart and is calculated by dividing the SV by the EDV. Note that the left ventricular EF (LVEF) is a measure of the efficiency of pumping blood into the body's systemic circulation, whereas the right ventricular EF (RVEF) is a measure of the efficiency of pumping blood into pulmonary circulation (i.e. the lungs).

- Left Ventricular Mass (LVM) (g): to compute LVM, we assume that the volume of the myocardium is equal to the total volume contained within the epicardial borders of the ventricle minus the chamber volume. Given these standard assumptions, LVM is calculated by multiplying the volume by the density of the muscle tissue (1.05 g/cm^3).

The regional assessment of cardiac function is based on the following indexes obtained from the LV myocardial shapes and computed locally based on the AHA 17-segment model. In contrast to the global indexes, where comparison with manual analysis was desired, and therefore 2D techniques were required (Simpson's rule), this segmental analysis was performed directly on 3D shapes, and using 3D techniques. Every measurement was computed on a per-point basis, and then averaged across all subjects, for every AHA-17 segment.

- LV Wall Thickness (mm): the distance between the endocardial and epicardial walls of the myocardium at ED and ES. Wall thickness may be used to quantify regional dysfunction, e.g. in myocardial ischaemia or after myocardial infarction. Myocardial thickness was measured as the average point-to-surface distance for every AHA-17 segment across the population.

- LV Wall Thickening (mm): the difference in the wall thickness measurement between ED and ES. Our models do not include papillary muscle or trabecular tissue, nor do the manual contours we compare our measurements with.

- LV Wall Motion (mm): the root-mean-squared distance between the location of mesh points at ED and ES averaged per AHA-17 region of the myocardium.

In the next section, we present and compare all the aforementioned global and regional clinical indexes obtained through manual and automatic segmentation.

3. Experiments and Results

We evaluated the performance of our proposed automated workflow by using common metrics for segmentation accuracy assessment (i.e. the aforementioned DSC, MCD and HD measures), comparing these measures against the ground-truth values obtained through manual delineation by clinicians and using clinical cardiac bi-ventricular functional indexes derived from manual and automated segmentations such as EDV, ESV and LVM.

We also compared our results with those reported by Bai et al. (2018). In Table 2, we present the data we used for training, testing and evaluating our workflow. Of the 4,870 available subjects in the UKB with manual segmentations, 250 random subjects were selected for PDM training, with 170 image volumes from a previous study by Tobon-Gomez et al. (2012) used for IAM training. The remaining 4,620 subjects in the UKB with manual delineations were used as test datasets to evaluate the performance of our proposed automatic approach, labelled $\mathrm{A_S}$ in the table. To compare our results with those of Bai et al. (2018), denoted B in the table, we used the same training and testing datasets, reporting the results as $\mathrm{A_L}$ in the table. As an additional assessment, we conducted a quantitative evaluation of human performance by measuring the inter-observer variability among the segmentations performed manually by three different clinical experts. Here, we randomly selected 50 subjects; each subject was independently analysed by three expert observers labelled O1, O2 and O3. We compare segmentation results on the same set of subjects to show automated versus human performance, as well as the performance of our workflow on a larger 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. included both basal and apical slices and to verify the quality of the output segmentations. Because our aim here is to properly evaluate segmentation accuracy, all segmentation results (including outliers) were included in the statistics in Section 3.1. In contrast, 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 providing full coverage by IQA.

Table 2: Specific datasets used for training and testing the methods proposed and presented in this paper.

Label	Method	Training/Tuning Data	Test Data
B	Method by Bai et al. (2018)	4,275 subjects from UKB	1) 600 subjects from UKB 2) 50 subjects from UKB
A_S	Our method (Small training dataset)	PDM: 250 subjects from UKB IAM: 170 subjects from Tobon-Gomez et al. (2012)	1) 4,620 subjects from UKB 2) 600 subjects from UKB 3) 50 subjects from UKB
A_L	Our method (Large training dataset)	PDM: 4,275 subjects from UKB IAM: 4,275 subjects from UKB	600 subjects from UKB
O1-O3	Human readers	Manual contours	50 subjects from UKB (three expert readers)

3.1. Segmentation Accuracy

To quantify segmentation accuracy, we applied the three aforementioned metrics, each of which is detailed below. First, the DSC evaluates the overlap between automated segmentation $\mathbf{A}$ and manual segmentation $\mathbf{M}$; we define DSC as follows:

$$DSC = \frac{2|\mathbf{A} \cap \mathbf{M}|}{|\mathbf{A}| + |\mathbf{M}|} \tag{7}$$

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

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

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

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

Table 3 presents DSC, MCD and HD measures that compare automated and manual segmentation results; evaluations were performed on test sets consisting of 50, 600 and 4,620 subjects which have not been used to train the PDM or IAM. Here, the group of 50 subjects is the same set used to evaluate inter-observer variability, whereas the set of 600 subjects is the same set used as a test set in Bai et al. (2018) in which a deep learning approach was used for segmentation. The large set of 4,620 subjects is 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} with $n = 4,620$ were 0.93 ± 0.05, 0.87 ± 0.05, and 0.87 ± 0.07, respectively, indicating excellent agreement between manual delineations and automated segmentations. We also observe that DSC measures for the LV_{myo} and RV_{endo} cases were less than that of the LV_{endo} case. One possible reason DSC values for the LV_{myo} are lower is that its annular shape has a larger perimeter (i.e. endo and epicardial edge) causing equal overlap shifts to produce greater error compared to the LV_{endo} and RV_{endo}.

Further, the RV is a more challenging structure to segment compared to the LV. This is due to the subpixel thickness of the RV myocardium, the larger 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 base to apex, and considerable variability in shape and intensity of the chamber across subjects, notably in pathological cases.

Next, we observe that 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 to 2.3 mm. The HD measures were 3.44 ± 1.08 mm, 3.98 ± 1.49 mm and 7.84 ± 3.19 mm for the LV_{endo}, LV_{myo} and RV_{endo}, respectively. Although HD measures are larger than the in-plane pixel spacing, they are still within acceptable range compared to inter-observer variability. For instance, the first three columns of Table 3 show inter-observer variability, where the variability between observers O1 and O2 for the HD metric is 7.56 ± 5.51 mm.

When comparing our method (i.e. A_S and A_L) with B, there was a notable difference in performance between the relatively small training set (i.e. A_S) and the same training set as that of B (i.e. A_L). In Table 3, we note a slight improvement of

Table 3: Segmentation results based on the different test sets (see Table 2) used in Bai et al. (2018), and Petersen et al. (2017) (n=50, 600, and 4,620). The metrics used are DSC, MCD and HD. We compare manual with automatic methods, and error between human observers. M represents the manual ground-truth provided by Petersen et al. (2017). $\mathrm{LV_{endo}}$ represents LV endocardium, $\mathrm{LV_{myo}}$ represents LV myocardium, and $\mathrm{RV_{endo}}$ represents RV endocardium. Table values are shown as mean ± standard deviation.

(a) DSC

	O1 vs O2	O2 vs O3	O3 vs O1	B vs M	$\mathrm{A_S}$ vs M	B vs M	$\mathrm{A_S}$ vs M	$\mathrm{A_L}$ vs M	$\mathrm{A_S}$ vs M
Test-set	(n=50)	(n=50)	(n=50)	(n=50)	(n=50)	(n=600)	(n=600)	(n=600)	(n=4620)
$\mathrm{LV_{endo}}$	0.94 ± 0.04	0.92 ± 0.04	0.93 ± 0.04	0.94 ± 0.04	0.93 ± 0.03	0.94 ± 0.04	0.93 ± 0.05	0.94 ± 0.04	0.93 ± 0.05
$\mathrm{LV_{myo}}$	0.88 ± 0.02	0.87 ± 0.03	0.88 ± 0.02	0.87 ± 0.03	0.88 ± 0.03	0.88 ± 0.03	0.87 ± 0.04	0.87 ± 0.03	0.87 ± 0.05
$\mathrm{RV_{endo}}$	0.87 ± 0.06	0.88 ± 0.05	0.89 ± 0.05	0.86 ± 0.07	0.87 ± 0.06	0.90 ± 0.05	0.88 ± 0.06	0.89 ± 0.05	0.87 ± 0.07

(b) MCD (mm)

	O1 vs O2	O2 vs O3	O3 vs O1	B vs M	$\mathrm{A_S}$ vs M	B vs M	$\mathrm{A_S}$ vs M	$\mathrm{A_L}$ vs M	$\mathrm{A_S}$ vs M
Test-set	(n=50)	(n=50)	(n=50)	(n=50)	(n=50)	(n=600)	(n=600)	(n=600)	(n=4620)
$\mathrm{LV_{endo}}$	1.00 ± 0.25	1.30 ± 0.37	1.21 ± 0.48	1.08 ± 0.30	1.28 ± 0.39	1.04 ± 0.35	1.21 ± 0.36	1.06 ± 0.35	1.18 ± 0.41
$\mathrm{LV_{myo}}$	1.16 ± 0.34	1.19 ± 0.25	1.21 ± 0.36	1.18 ± 0.31	1.20 ± 0.34	1.14 ± 0.40	1.23 ± 0.48	1.13 ± 0.35	1.23 ± 0.50
$\mathrm{RV_{endo}}$	2.00 ± 0.79	1.78 ± 0.45	1.87 ± 0.74	2.20 ± 0.92	1.79 ± 0.80	1.78 ± 0.70	1.80 ± 0.80	1.74 ± 0.61	1.80 ± 0.69

(c) HD (mm)

	O1 vs O2	O2 vs O3	O3 vs O1	B vs M	$\mathrm{A_S}$ vs M	B vs M	$\mathrm{A_S}$ vs M	$\mathrm{A_L}$ vs M	$\mathrm{A_S}$ vs M
Test-set	(n=50)	(n=50)	(n=50)	(n=50)	(n=50)	(n=600)	(n=600)	(n=600)	(n=4620)
$\mathrm{LV_{endo}}$	2.84 ± 0.70	3.31 ± 0.90	3.25 ± 0.96	3.46 ± 1.05	3.21 ± 0.97	3.16 ± 0.98	3.29 ± 1.04	3.15 ± 0.96	3.44 ± 1.08
$\mathrm{LV_{myo}}$	3.70 ± 1.16	3.82 ± 1.07	3.76 ± 1.21	4.06 ± 1.16	3.91 ± 1.20	3.92 ± 1.37	3.97 ± 1.43	3.90 ± 1.29	3.98 ± 1.49
$\mathrm{RV_{endo}}$	7.56 ± 5.51	7.35 ± 2.19	7.14 ± 2.20	9.02 ± 3.54	7.41 ± 4.11	7.25 ± 2.70	7.54 ± 3.20	7.21 ± 2.62	7.84 ± 3.19

the mean and standard deviation values, particularly for MCD measures. Nevertheless, improvements become more apparent in Figure 3, where the number of outlying subjects was drastically reduced for $\mathrm{A_L}$ as compared to both B and $\mathrm{A_S}$. Although the overall mean and standard deviation values remained slightly better for B, we observe in Figure 3 that $\mathrm{A_L}$ was generally more robust as it reduced the number and deviation of outlying results.

Also from Figure 3, we note that the performance of $\mathrm{A_S}$ largely agrees with the ground-truth and is comparable to the results of B. We also investigated the segmentation accuracy of the LV myocardium in detail based on the AHA 17-segment model of Heller et al. (2002) to report on local segmentation accuracy in terms of DSC, MCD and HD measures between manual segmentation and automatic approaches, i.e. B and $\mathrm{A_S}$ on the test set of size 600. We report local segmentation accuracy in Table 4, which shows that B and $\mathrm{A_S}$ consistently performed better with mid-ventricular and apical slices, respectively; however, for base slices, the performance of B and $\mathrm{A_S}$ varies per region.

Note that when comparing the performance of $\mathrm{A_S}$ versus B ($n = 600$) in Table 3, B yielded slightly better global results than $\mathrm{A_S}$, but in breaking down the results into specific cardiac regions (basal, mid and apical), as presented in Table 4 we observe that our method, $\mathrm{A_S}$, consistently outperformed B for all metrics in the apical region (AHA segments 13-17). A possible reason for this is an inability of the CNN method to capture small features in the image, and the inherent ability of PDMs to infer missing or noisy image data.

To provide a visual sense of the quality of our segmentations, we defined three categories based on the mean contour distance from the gold standard, i.e., excellent (MCD < 1 pixel), good (1 pixel < MCD < 2 pixels) and bad (MCD > 2 pixels). We present examples of these categories in Figure 4, thereby showing that automated segmentation agrees well with manual segmentation both at ED and ES; further, such agreement occurs at different slice locations (i.e. apical, mid and basal regions). Finally, Table 5 shows the prevalence of the different categories of segmentation quality for the different approaches presented in this paper.

3.2. Estimation of Cardiac Function Indexes

In this subsection, we present our work in evaluating the accuracy of cardiac function indexes derived from automated segmentation using gold standard reference ranges derived from manual segmentations. Further, we report on analysis of all available CMR images from the UKB, which to date is 20K subjects. More specifically, we calculate the

Table 4: Regional segmentation accuracy of LV myocardium based on the AHA 17-segment model covering 600 subjects. Values indicate mean ± standard deviation. **Bold** indicates the cases in which our algorithm (i.e. A_S) outperformed algorithm B.

		DSC		MCD		HD	
		B vs M	A_S vs M	B vs M	A_S vs M	B vs M	A_S 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**

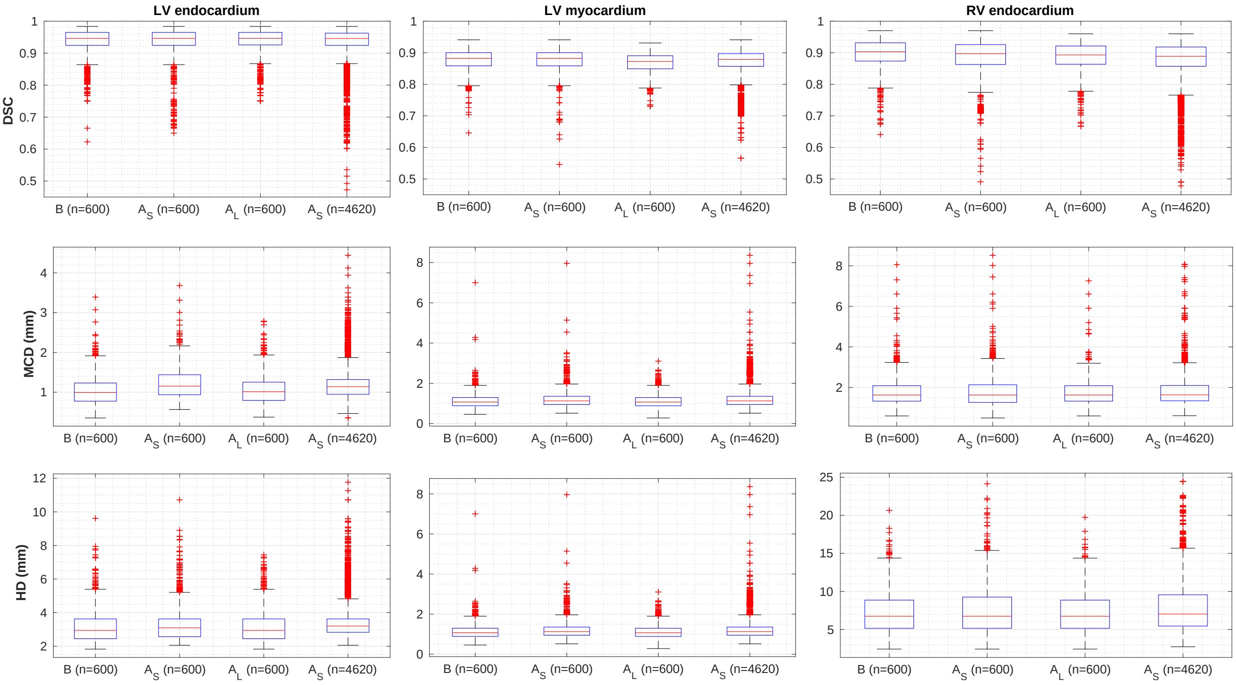


Figure 3: Segmentation accuracy expressed in terms of the DSC, MCD and HD measures.

following two sets of indexes: (1) *global indexes* including the LV end-diastolic volume (LVEDV) and end-systolic volume (LVESV), LV stroke volume (LVSV), LV ejection fraction (LVEF), LV myocardial mass (LVM), RV end-diastolic volume (RVEDV) and end-systolic volume (RVESV), RV

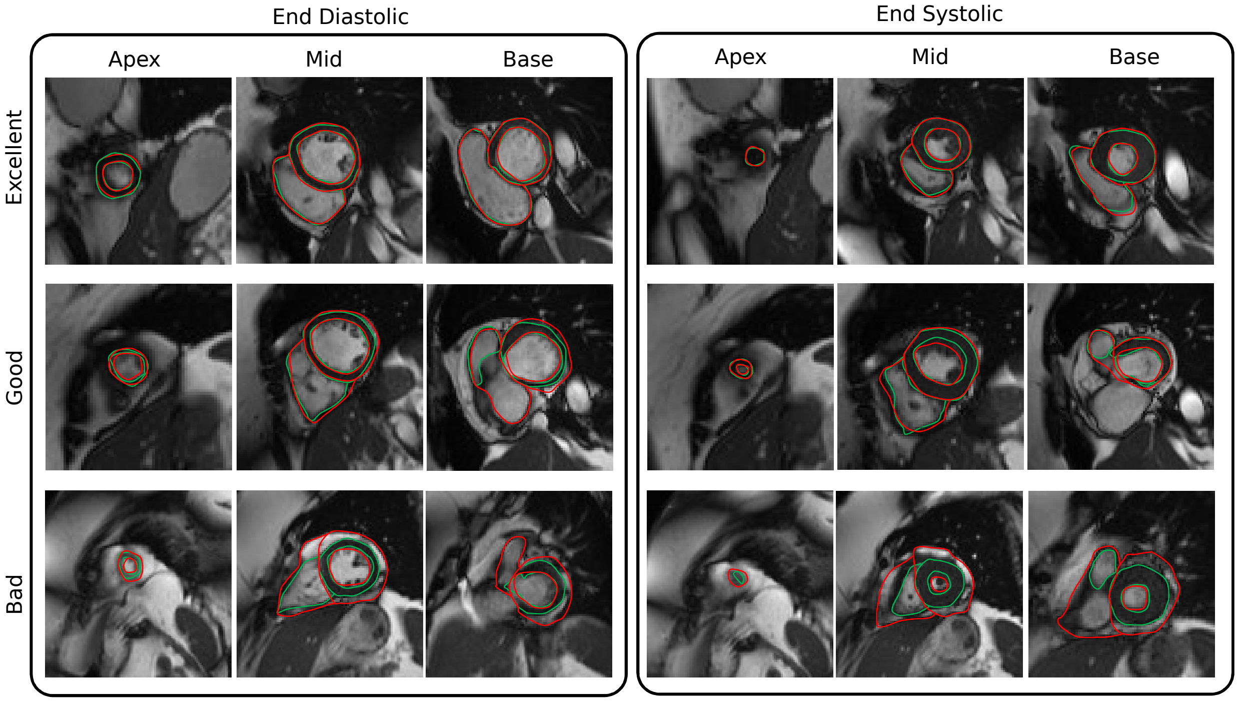


Figure 4: Examples segmentation results at the ED and ES phases illustrating our three degrees of quality for automated segmentation contours versus manual contours. Red: automated segmentations. Green: ground-truth segmentations.

Table 5: Categories of segmentation quality for the different approaches presented in this paper.

	B (n=600)	A_S (n=600)	A_L (n=600)	A_S (n=4620)
Excellent MCD <1 pixel	84.19 %	82.14 %	84.21 %	80.67 %
Good 1 pixel <MCD <2 pixels	15.25 %	16.80 %	15.50 %	17.50 %
Bad MCD >2 pixels	0.55 %	1.05 %	0.30 %	1.82 %

stroke volume (RVSV) and RV ejection fraction (RVEF); and (2) *regional indexes* including the myocardium wall thickness, thickening and motion.

Note that we report the clinical indexes obtained from automated segmentation of subjects that have successfully passed the IQA and SQA modules. Table 6 shows the number of subjects that were included in our analysis. For example, of the given 4,620 subjects, 4,430 were deemed of good quality after IQA and SQA analyses were applied. More specifically, IQA detected 145 subjects to exclude, whereas SQA detected 105 subjects to omit; note that 60 subjects were common to both lists. Therefore, a total of 190 subjects were automatically removed before continuing with the analysis.

Table 7 shows the main cardiac clinical indexes, with the two first columns representing the ventricular parameters of the healthy population obtained through automated and manual segmentations. We observe here that there was strong agreement between the two methods for computing the presented cardiac function indexes (Attar et al., 2018). Similarly, the computed clinical indexes for the large cohort of 4,620 subjects correlated well with the corresponding ground-truth values, as shown in columns three and four of the table; however, we note that although the mean and standard deviation values of the RV indexes for the healthy population of 800

Table 6: A summary of subjects used in our analysis after quality control measures were applied.

Datasets	Total number of subjects (n)	Detected by IQA only	Detected by SQA only	Detected by IQA & SQA	Remain for analysis
Healthy population (Petersen et al., 2017)	800	0	21	0	779
All manually segmented	4,620	145	105	60	4,430
Dataset used in (Bai et al., 2018)	600	0	11	0	589
UKB dataset	20,000	284	234	138	19,620

Table 7: Summarising the differences in clinical measures derived from our proposed method and manual segmentation. Here, GT represents the ground-truth values provided by manual segmentation from Petersen et al. (2017). Values indicate mean ± standard deviation.

	GT (n=800)	Automated (n=800)	GT (n=4,620)	Automated (n=4,620)	Automated (n=20,000)
LVEDV (ml)	144 ± 34	**146 ± 31**	144 ± 34	**144 ± 33**	**142 ± 26**
LVESV (ml)	59 ± 18	**60 ± 18**	59 ± 20	**60 ± 23**	**53 ± 14**
LVSV (ml)	85 ± 20	**86 ± 18**	84 ± 18	**84 ± 19**	**89 ± 18**
LVEF (%)	60 ± 6	**60 ± 7**	60 ± 6	**59 ± 7**	**63 ± 6**
LVM (g)	86 ± 24	**87 ± 23**	88 ± 23	**91 ± 23**	**92 ± 18**
RVEDV (ml)	154 ± 40	**154 ± 40**	152 ± 37	**160 ± 49**	**165 ± 41**
RVESV (ml)	69 ± 24	**71 ± 26**	67 ± 22	**77 ± 26**	**61 ± 24**
RVSV (ml)	85 ± 20	**83 ± 21**	84 ± 18	**82 ± 24**	**90 ± 27**
RVEF (%)	56 ± 6	**54 ± 7**	57 ± 6	**54 ± 11**	**60 ± 9**

subjects were in good agreement, for the population of 4,620 subjects, the mean and standard deviation values of the RV indexes differed slightly compared with the ground-truth values. This correlates with the larger inter-observer variability shown in Table 3, which is at least in part due to thinness of the RV myocardium vis-a-vis the LV (Zheng et al., 2018).

Table 8 presents the mean absolute and relative differences between the automated and manual measurements, as well as between the automated and manual measurements computed by different expert human observers and by the built-in automated segmentation software of the scanner device (i.e. inlineVF D13A). We observe here that the absolute and relative differences for two subsets of 50 and 600 subjects matched well and were within the error range of the three expert human observers. Similarly, although the range of differences over the cohort of 4,620 subjects were not directly comparable with a small test set of only 50 subjects, the difference range still was either within that range or very close to the difference range obtained by the different expert observers. Overall, B, $\mathrm{A_S}$ and $\mathrm{A_L}$ performed substantially better than the automated segmentation obtained from the inlineVF D13A software; note that these data were retrieved for every subject from the main UKB database.

Next, in Figure 5, we present Bland-Altman plots (i.e. the top row of the figure) and correlation plots (i.e. the bottom row of the figure) of the ventricular parameters computed based on our proposed automated method and a manual reference covering 4,620 test subjects. The Bland-Altman plot is commonly used for analysing agreement and bias between two measurements. In Figure 5, the Bland-Altman plots show strong agreement and a mean difference line at nearly zero, suggesting that the clinical indexes obtained through the automated approach have little bias. Conversely, the bias between different pairs of human observers as reported by Bai et al. (2018) is considerable – i.e. nearly 8 (ml) for LVEDV and LVESV, approximately 8 (g) for LVM, and approximately 15 (ml) for RVEDV and RVESV.

More specifically, Figure 5 presents correlation plots between the manual and automated methods for the different cardiac function indexes. The correlation coefficient (corr) measures 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 ranged between 0.85 and

Table 8: 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 (Petersen et al., 2017). VF: Automatic segmentation obtained from the automatic segmentation software inlineVF D13A. Values indicate mean ± standard deviation.

(a) Absolute difference

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

(b) Relative difference (%)

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

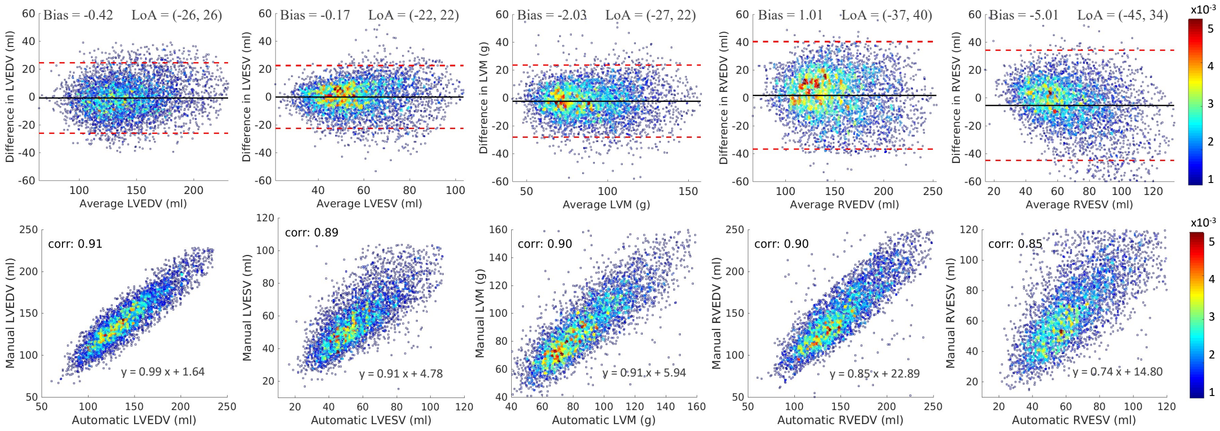


Figure 5: Illustrating the repeatability of various cardiac functional indexes comparing the manual and automated analysis of 4,620 subjects from the UKB cohort. The top row shows Bland-Altman plots for various cardiac functional indexes computed both manually and automatically in which manual segmentation was available. The black horizontal lines denote the mean difference (i.e. bias), whereas the two red dashed lines denote limits of agreement (LoA) i.e. ±1.96 standard deviations from the mean. The second row shows correlation plots for various cardiac functional indexes computed both manually and automatically in which manual segmentation was available.

0.91, indicating a strong relationship between the manual and automated approaches.

To illustrate whether the values of clinical indexes computed automatically share the same distribution as those obtained via the manual approach, we visualised their distributions. In Figure 6, we present probability distribution plots (i.e. the top row of the figure) and Q-Q plots (i.e. the bottom row of the figure) for various cardiac functional indexes computed both manually and automatically over the full cohort for which manual segmentations were available. From the plots, we observe that the distribution of the various indexes closely match those obtained from the manual segmentations– More specifically, we observe a common distribution, common location and scale, similar distributional shapes, and similar tail behaviour.

Because ground-truth manual regional (AHA-17) quantification for the subjects in this study was not available, all AHA-17 regional indexes reported in this paper are computed using 3D techniques, in contrast to the global quantification indexes, where direct comparison with manual assessment was desirable. Nevertheless, in order to approximate a comparison with what would be a regional analysis derived from manual delineations, we generated

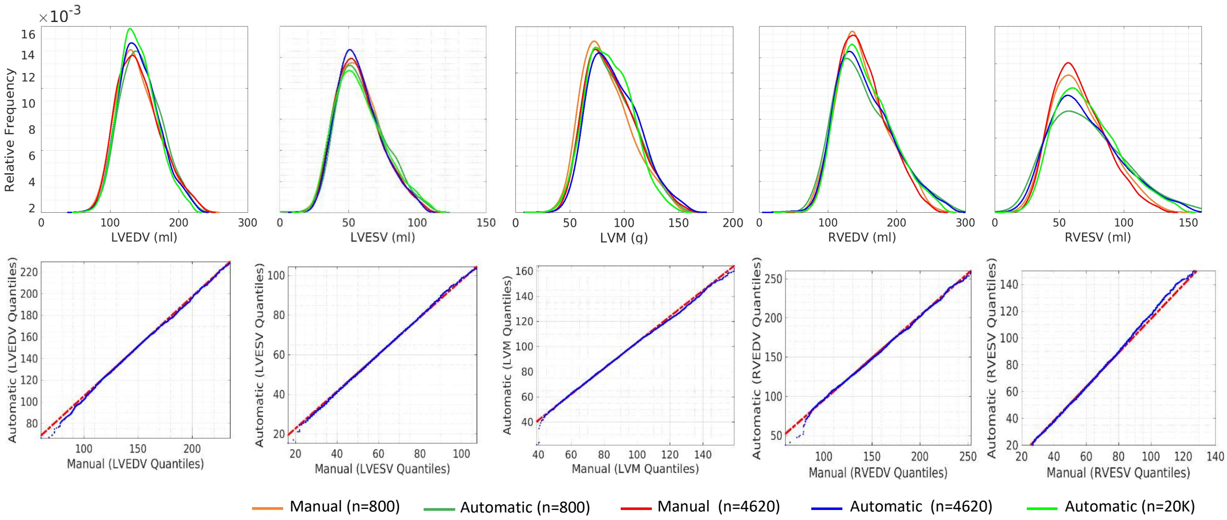


Figure 6: Distributions of various cardiac functional indexes comparing results of manual and automatic analyses of 4,620 subjects. The top row shows probability distribution plots, whereas the bottom row shows Q-Q plots for various cardiac functional indexes computed both manually and automatically in which manual segmentation was available.

3D shapes by non-rigid registration of a model to all manual delineations. We used the resulting 3D shapes to perform regional quantification, and compared with our automatic results.

We computed the regional LV myocardial wall parameters in terms of thickness, thickening, and motion. Visual results can be seen on Figure 7, and corresponding numerical results on Table 9. Figure 7 shows the mean and standard deviation values of the regional analysis of 4,620 subjects for both the automated and manual approaches in a bulls-eye display based on the AHA 17-segment model. We observe here that the (top and bottom) panels are similar in most regions in terms of the mean and standard deviation values, thereby confirming the quality of our fully automated pipeline. Indeed, results already published in many clinical journals (Andre et al., 2012; Deviggiano et al., 2016; Puntmann et al., 2010; Kanza et al., 2007; Prasad et al., 2010; Le Ven et al., 2015; Baltabaeva et al., 2007; Codreanu et al., 2014), primarily based on the manual delineation of a few dozen images confirm the values and ranges we have obtained and present in our bulls-eye plots.

Figures 8 and 9 show the distribution of wall thickness, thickening and motion for all AHA-17 segments in the LV myocardium. These histograms show measurements obtained from the automated segmentation applied to two cohorts (i.e. n=4,620 and n=20,000), as well as from manual delineations. The figures show excellent agreement between measurements obtained from automated segmentations from both cohorts and those derived from manual delineations.

We also performed two-sample Kolmogorov-Smirnov (K-S) tests to verify that ventricular parameters obtained through manual and automated approaches are drawn from the same distribution, under the null hypothesis that the manual and automatic methods are from the same continuous distribution in terms of clinical indexes. From our analysis, K-S test results on different global and regional indexes do not reject the null hypothesis of being from the same distribution at the 5% significance level.

An important final note is that although our image parsing implementation performs fully in 3D, to ensure a fair comparison with both ground-truth data and the methods we compare with in this paper, we had to convert our segmentation results to 2D contours from 3D meshes; this does not pose a problem for objective quantification of segmentation accuracy, however, given the sparse nature of CMR images, where voxel resolution along the z axis is typically on the order of $10mm$, gross miscalculations may occur when approximating volumetric measurements such as ventricular volumes and myocardial masses via simple integration methods such as Simpson's rule. We believe that although many CNN-based methods have recently received a

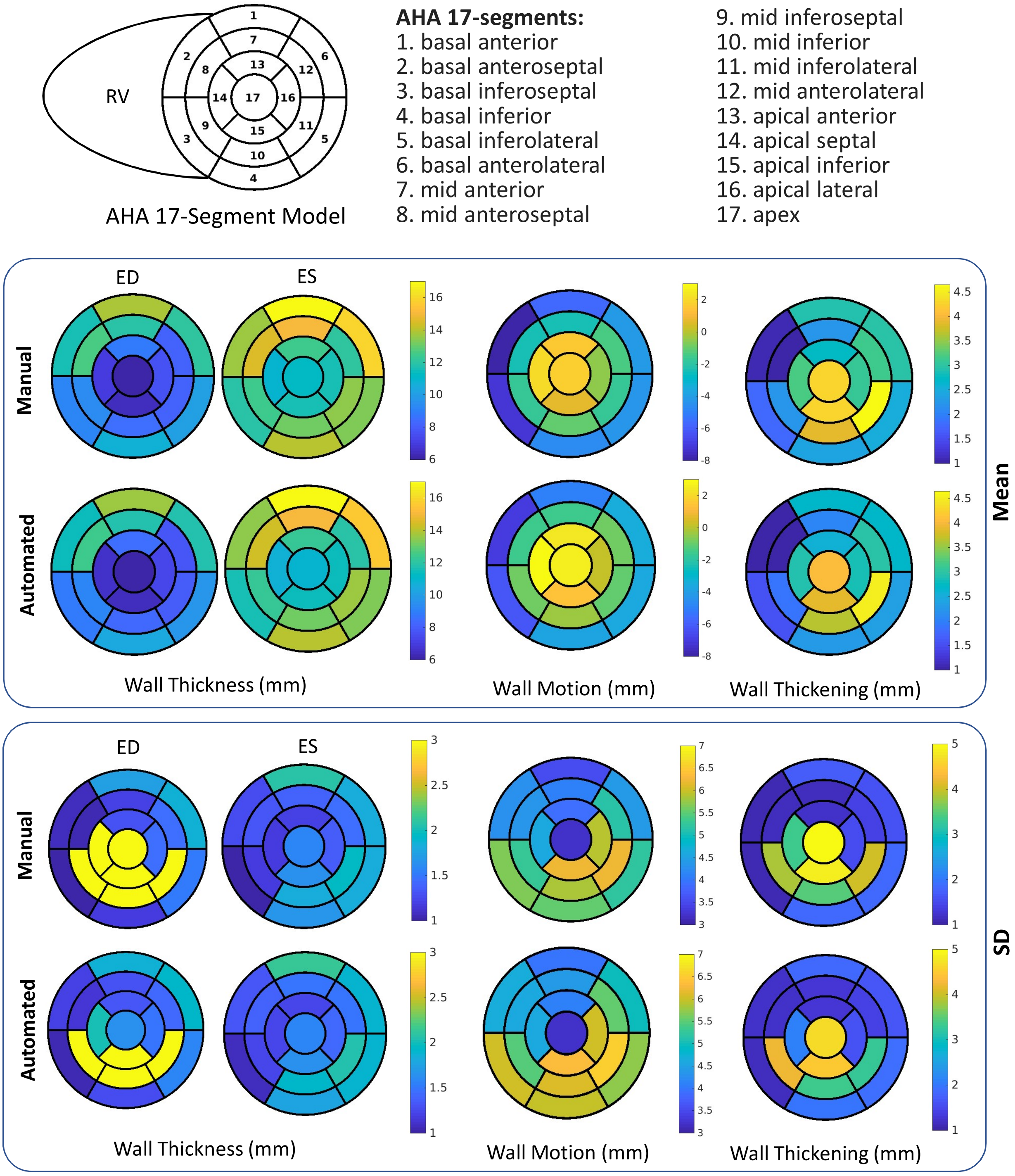


Figure 7: Segmental LV parameters of 4,620 subjects presented as bulls-eye displays.

lot of attention, showing the capacity for image texture characterisation, most of them are restricted to handling 2D data. This simplification can introduce large biases in volume computations, and be less resilient to image artefacts such as those caused by breathing motion. In addition to our pipeline ap-

Table 9: Segmental LV parameters of 4,620 subjects obtained from manual and automatic approaches. Upper rows correspond to shapes generated from the manual segmentation and lower rows to those obtained with the automatic approach.

ID	Wall Thickness at ED (mm)	Wall Thickness at ES (mm)	Wall Motion (mm)	Wall Thickening (mm)
1	12.05 ± 1.84	15.69 ± 1.83	-4.05 ± 4.39	2.95 ± 1.61
	11.72 ± 1.92	15.59 ± 1.88	-3.46 ± 4.94	2.67 ± 1.66
2	13.81 ± 1.73	17.49 ± 2.07	-5.70 ± 3.51	2.89 ± 1.74
	13.59 ± 1.85	17.49 ± 2.14	-4.91 ± 3.97	2.69 ± 1.81
3	11.63 ± 1.09	13.57 ± 1.26	-8.01 ± 4.28	1.08 ± 1.10
	11.38 ± 1.20	13.47 ± 1.33	-7.06 ± 4.63	0.87 ± 1.19
4	9.17 ± 0.97	11.87 ± 1.00	-7.20 ± 5.68	1.76 ± 1.08
	8.87 ± 1.06	11.65 ± 1.06	-6.20 ± 6.03	1.55 ± 1.18
5	10.76 ± 1.17	14.00 ± 1.67	-4.60 ± 5.53	2.41 ± 1.86
	10.45 ± 1.27	13.96 ± 1.77	-3.80 ± 5.96	2.29 ± 1.95
6	10.02 ± 1.50	13.35 ± 1.81	-4.29 ± 5.21	2.52 ± 1.82
	9.69 ± 1.58	13.26 ± 1.89	-3.60 ± 5.72	2.35 ± 1.87
7	8.14 ± 1.46	12.14 ± 1.44	-1.44 ± 5.18	3.15 ± 1.41
	7.81 ± 1.43	11.96 ± 1.48	-1.31 ± 5.50	2.94 ± 1.42
8	11.99 ± 1.21	15.03 ± 1.40	-2.33 ± 4.06	2.25 ± 1.18
	11.68 ± 1.31	14.95 ± 1.46	-1.57 ± 4.40	2.04 ± 1.26
9	12.63 ± 1.04	14.51 ± 1.32	-2.35 ± 4.37	1.01 ± 1.12
	12.37 ± 1.15	14.41 ± 1.40	-1.52 ± 4.62	0.84 ± 1.20
10	9.33 ± 3.59	12.43 ± 1.17	-1.78 ± 5.26	2.25 ± 3.83
	9.17 ± 4.16	12.28 ± 1.24	-0.93 ± 5.46	1.88 ± 4.23
11	8.53 ± 3.12	13.09 ± 1.82	-1.02 ± 5.85	3.85 ± 3.49
	8.25 ± 3.07	13.11 ± 1.91	-0.84 ± 6.11	3.66 ± 3.33
12	8.17 ± 3.76	13.55 ± 1.94	-1.72 ± 6.23	4.65 ± 4.00
	7.90 ± 3.14	13.59 ± 2.03	-1.09 ± 6.89	4.50 ± 3.30
13	7.58 ± 1.45	11.57 ± 1.37	-0.58 ± 5.91	3.10 ± 1.50
	7.32 ± 1.49	11.40 ± 1.42	-0.39 ± 6.05	2.86 ± 1.56
14	8.75 ± 1.30	12.40 ± 1.19	1.56 ± 3.91	2.75 ± 1.22
	8.46 ± 1.33	12.21 ± 1.23	2.32 ± 4.10	2.54 ± 1.28
15	6.86 ± 3.17	10.89 ± 1.17	2.07 ± 4.59	3.20 ± 3.36
	6.59 ± 2.02	10.70 ± 1.21	2.93 ± 4.44	2.91 ± 2.09
16	6.48 ± 4.66	11.58 ± 1.63	0.69 ± 6.22	4.29 ± 4.83
	6.37 ± 4.44	11.46 ± 1.63	0.41 ± 6.68	3.90 ± 4.49
17	5.97 ± 4.49	11.17 ± 1.57	1.62 ± 3.21	4.28 ± 5.11
	5.71 ± 3.63	10.94 ± 4.68	2.58 ± 3.18	4.03 ± 1.68

proach fully supporting 3D data, our method provides other advantages when compared to 2D CNN-based implementations. More specifically, the size of the training dataset required to achieve similar performance for an equal task differs by at least one order of magnitude between CNNs and ASM-based methods. Further, ASM implementations such as SPASM have the inherent ability to handle multi-view image volume segmentation without the need to retrain. This is particularly useful for functional CMR segmentation in which multiple views of the heart are captured as part of standard analysis protocols. In addition, because the output of our segmentation are 3D meshes, more apt mathematical formulations can be used for volumetric computation, i.e. Green's theorem for surface integration, and any further higher level structural analyses of the cardiac tissue. Some CNN-based methods such as those proposed by Zheng et al. (2018) do take into account inter-dependencies between short-axis slices potentially resulting in more robust segmentations, even so, such CNN-based algorithms are still not globally constrained, their output is typically two dimensional in nature, their training is very costly both in time and sample size requirements, and they cannot handle dynamically changing input image views without redefinition of the architecture and re-training. We present the key

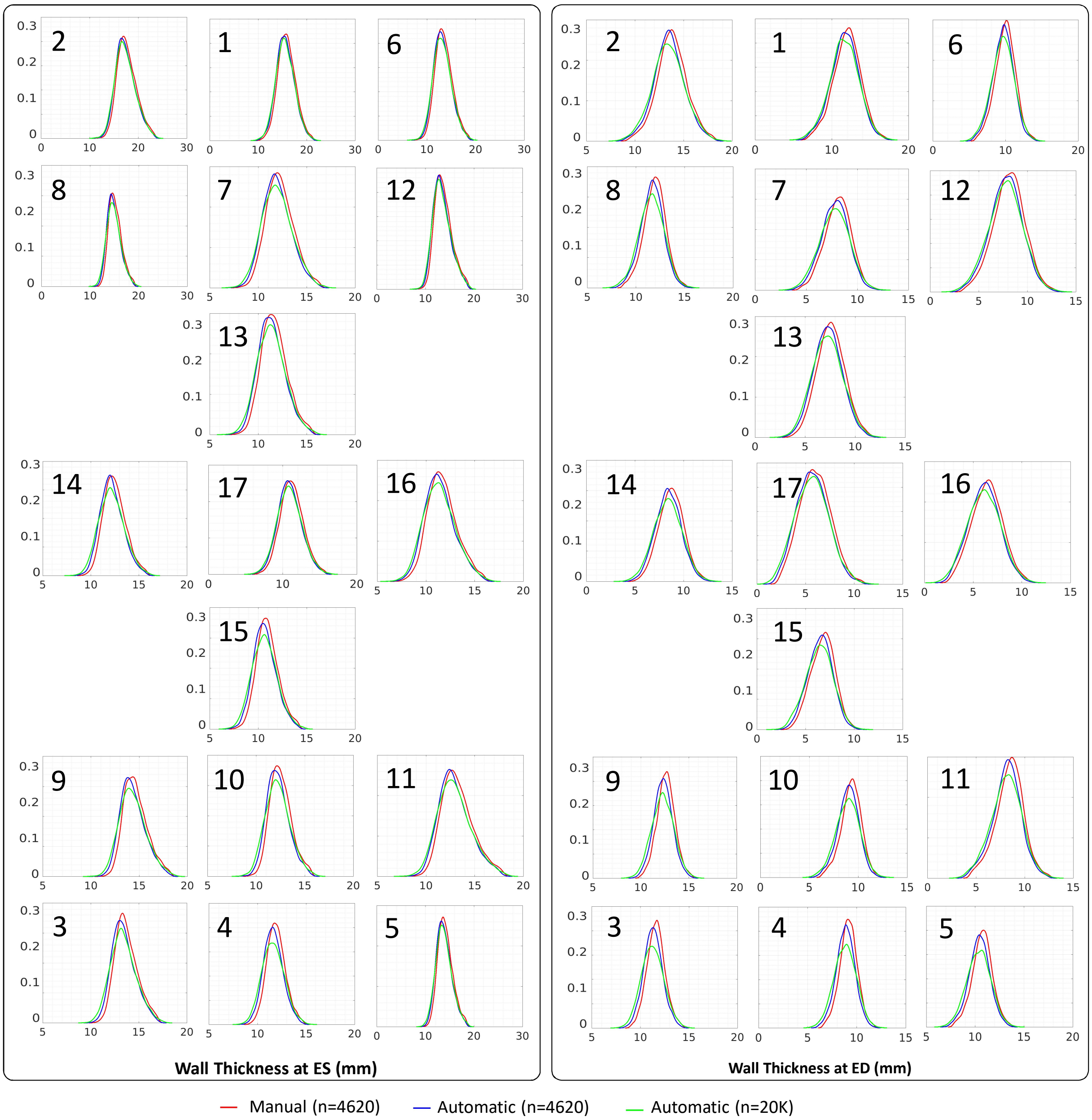


Figure 8: Regional analysis of LV shapes covering 20,000 subjects in terms of distribution of wall thickness at ED and ES phases. Here, red, blue and green lines indicate ground-truth values for 4,620 subjects, automated values for 4,620 subjects and automated values for 20,000 subjects, respectively. In all plots, the y-axis represents the relative frequency.

differences between our implementation and the 2D CNN-based implementation method by Bai et al. (2018) on Table 10.

3.3. Hardware and Computational Cost

In terms of computational cost of training and testing, method B takes approximately 10 hours to train the VGG-16 network on a Nvidia Tesla K80 GPU, and about 11 seconds to segment all 2D slices of a full cardiac cycle for one subject (Bai et al., 2018). For our method, it takes approximately 30 minutes to train both the PDM and IAM on a Intel Xeon(R) CPU E5-1620 @3.60GHz with 32 GB of RAM, and about 15 minutes to generate the 3D

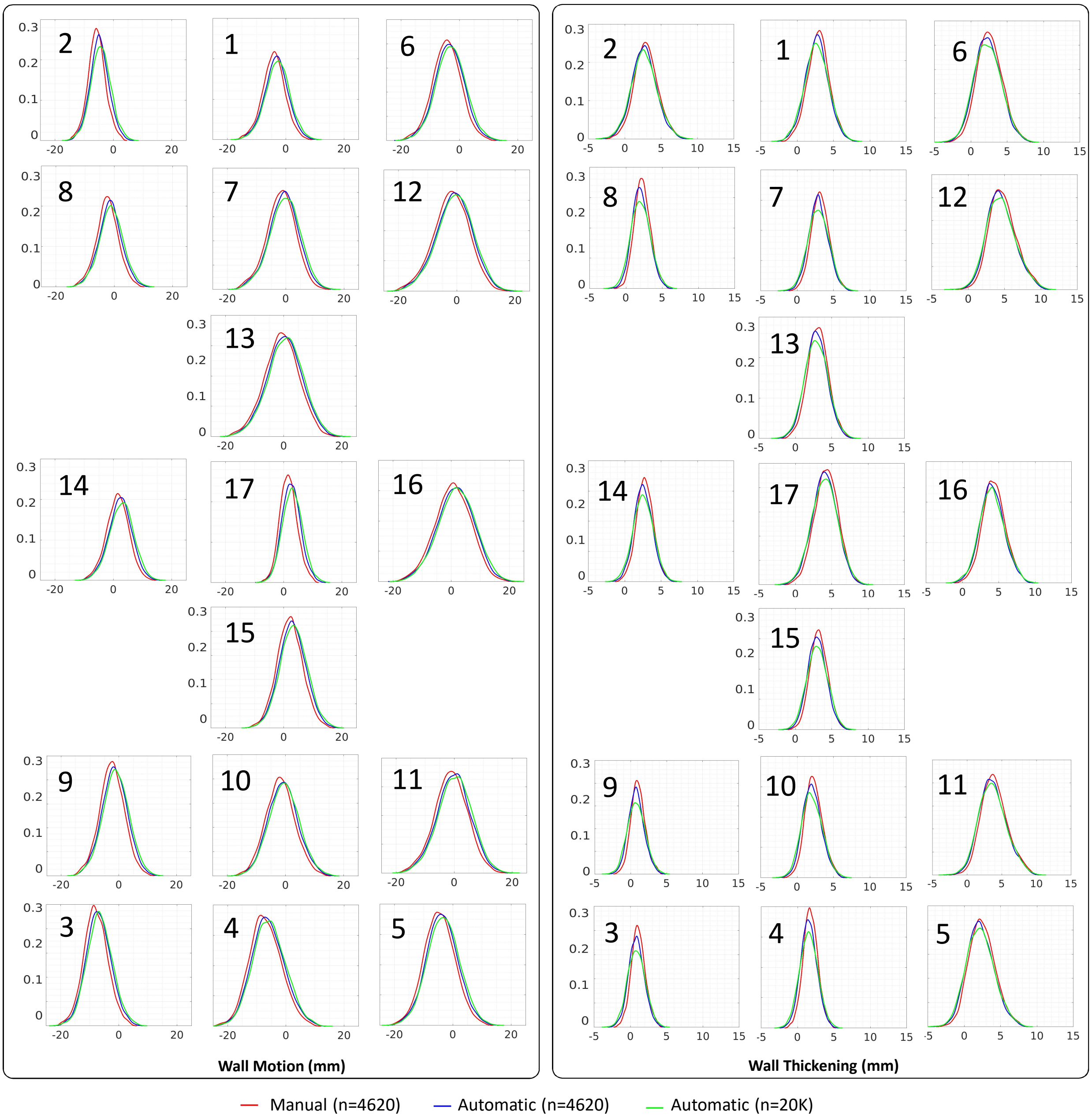


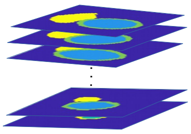
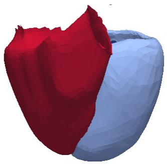
Figure 9: Regional analysis of LV shapes covering 20,000 subjects in terms of distribution of wall motion and thickening. Here, red, blue and green lines indicate ground-truth values for 4,620 subjects, automated values for 4,620 subjects and automated values for 20,000 subjects, respectively. In all plots, the y-axis represents the relative frequency.

shapes of a full cardiac cycle for one subject. Finally, the total end-to-end execution time for the 20,000 subjects using our MULTI-X platform was performed using 50 Amazon Web Service (AWS) ”m4.10xlarge” machines each with 40 2.4-GHz Intel Xeon ES-2676 v3 vCPUs, and 160 GB of RAM.

3.4. Sub-Cohort Analysis

Thus far in this paper we have only shown a global population analysis of the UKB. We have presented statistics on the most commonly used clinical indexes derived from CMR exams. With the exception of the ”healthy” group as defined by Petersen et al. (2017), introduced in this paper on

Table 10: Comparing one of the current state-of-the-art CNN-based methods proposed by Bai et al. (2018) with our proposed framework.

	Bai et al. (2018)	Proposed pipeline
Output	 2D masks	 3D surface mesh
Size of training dataset (N) for equal performance	$N \approx a \times 10^3$	$N \approx a \times 10^2$
Image slice dimensions	Must match training (cropping or re-sampling required)	Independent to image size
Image view (SAX, LAX)	Must be consistent, i.e. SAX	Independent to image view
Slice stack inter-dependency	Each slice processed independently	Slice stack handled by PDM
Training computational cost	High*	Low*
Testing computational cost	Low*	High*

*For more details see section 3.3.

Table 6, and corresponding quantification results shown on the first two columns of Table 7, we have only presented global population statistics. We believe however that the power of population studies lies in the opportunity to define and characterise human sub-populations.

Though in this paper our principal aim is to present the first fully automatic large-scale, global and segmental, 3D analysis of this magnitude we have included some preliminary quantification results on UKB sub-populations in this section. Based on the 20,000 subjects available, we have used patient age at the time of imaging, and patient gender (male, female), to present cardiovascular index reference ranges for these cohorts. Table 11 presents the arithmetic mean, and upper/lower bounds of the 95% prediction interval for each clinical index, and each age group. Each of the three age-groups span a 10-year interval, and the total age range includes patients 45 to 74 years old. Also, for each clinical index, and age-group we compute separate statistics for males and females.

Figure 10 shows the mean value for each of the five clinical indexes, for the three different age groups, and for males and females. Perhaps the most evident, and in some ways expected feature of these plots, is the consistent decline in cardiac volumes and cardiac mass with ageing. For the five indexes LVEDV, LVESV, LVM, RVEDV and RVESV, we see a decline of 9%, 15%, 7%, 8%, and 13% for males, and 11%, 17%, 5%, 6%, and 11% for females. As stated before, a deep analysis of sub-populations is out of the scope of this paper, nevertheless, we hope to have shown the potential of the techniques presented in this paper to gain insight from large population imaging studies.

4. Conclusions

In this study, we presented a fully automatic framework capable of performing high-throughput end-to-end 3D cardiac image analysis of 20,000 subjects. We validated our workflow on a reference cohort of 4,620 subjects for which both manual delineations and reference functional indexes exist. Our results show that differences between our automatic workflow and the manually obtained global and regional reference indexes are within the expected variability observed in human raters. As future work, we would like to increase the robustness of our pipeline to handle severe pathological morphology and variable image quality. We foresee including feed-back loops in our pipeline that would allow the automatic adjustment of segmentation parameters for image re-processing, based on our quality assessment modules. Such feed-back loops may include triggering of new modules designed to handle poor image quality, or imputation

Table 11: Male (M) and Female (F) ventricular reference ranges detailing mean, lower reference limit and upper reference limit by age group. Reference limits are derived by the upper and lower bounds of the 95% prediction interval for each parameter at each age group.

Age groups (years)		**45-54**			**55-64**			**65-74**		
Number of subjects		**3510**			**7408**			**8702**		
Male gender (%)		**43%**			**43%**			**52%**		
		lower	mean	upper	lower	mean	upper	lower	mean	upper
LVEDV (ml)	M	109	170	231	102	163	223	94	154	213
	F	95	140	184	88	132	175	80	124	168
LVESV (ml)	M	31	81	130	28	75	122	25	69	113
	F	25	58	91	21	53	85	15	48	82
LVM (g)	M	71	112	152	69	109	148	66	104	142
	F	44	78	111	43	76	108	42	74	107
RVEDV (ml)	M	115	197	279	112	191	269	105	182	259
	F	73	145	218	72	141	210	72	136	200
RVESV (ml)	M	38	102	144	34	96	137	31	89	127
	F	16	66	116	15	63	110	14	59	105

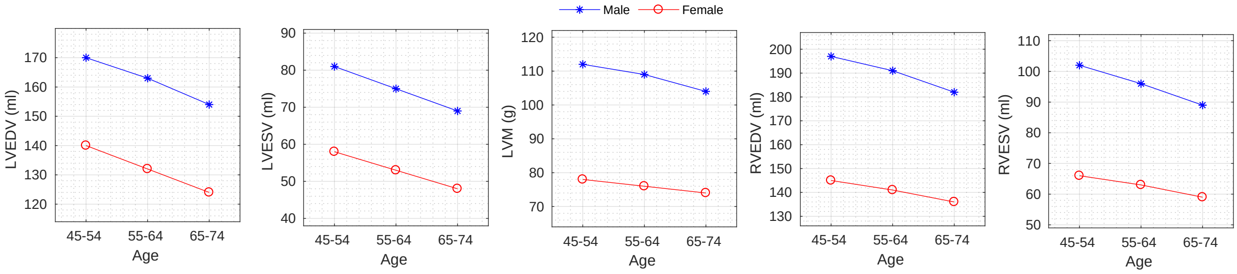


Figure 10: Male (blue star marker) and Female (red circle marker) clinical indexes showing their mean value per age group.

of missing data. Similarly, alternative segmentation techniques could be triggered upon detection of specific pathologies. Besides increasing the robustness of our system, we would like to further the analysis of reference ranges for specific sub-populations. The UKB provides a wealth of patient information including, socio-demographic, lifestyle and environmental, family history, genetic, and omics data. Modelling the relationship between these factors and cardiac morphology and function would help further our understanding of disease processes, and potentially increase the specificity of medical treatment.

5. Acknowledgements

RA was funded by the Faculty of Engineering Doctoral Academy Scholarship, University of Sheffield and School of Computing PhD Scholarship, University of Leeds. MULTI-X (www.multi-x.org) was partially supported by the European Commission through FP7 contract VPH-DARE@IT (FP7-ICT-2011-9-601055) and H2020 contract InSilc (H2020-SC1-2017-CNECT-2-777119). AFF acknowledges support from the Royal Academy of Engineering Chair in Emerging Technologies Scheme (CiET18199) and the MedIAN Network (EP/N026993/1) funded by the Engineering and Physical Sciences Research Council (EPSRC). SN acknowledges the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at The Oxford University Hospitals Trust at the University of Oxford, and the British Heart Foundation Centre of Research Excellence. AL and SEP acknowledge support from the NIHR Barts Biomedical Research Centre and from the SmartHeart" EPSRC program grant (EP/P001009/1). NA is supported by a Wellcome Trust Research Training Fellowship (203553/Z/Z). The UKB CMR dataset has been provided under UK Biobank Access Applications #2964 and #11350. The authors thank all UK Biobank participants and staff. Manual annotations of the CMR data was kindly provided by British Heart Foundation (PG/14/89/31194).

References

Albà, X., Lekadir, K., Pereañez, M., Medrano-Gracia, P., Young, A. A., and Frangi, A. F. (2018). Automatic initialization and quality control of large-scale cardiac mri segmentations. *Medical image analysis*, 43:129–141.

Andre, F., Lehrke, S., Katus, H. A., and Steen, H. (2012). Reference values for the left ventricular wall thickness in cardiac MRI in a modified AHA 17-segment model. *Journal of Cardiovascular Magnetic Resonance*, 14(S1):P223.

Attar, R., Pereañez, M., Gooya, A., Albà, X., Zhang, L., Piechnik, S. K., Neubauer, S., Petersen, S. E., and Frangi, A. F. (2018). High throughput computation of reference ranges of biventricular cardiac function on the uk biobank population cohort. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 114–121. Springer.

Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A. M., Aung, N., Lukaschuk, E., Sanghvi, M. M., et al. (2018). Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20(1):65.

Baltabaeva, A., Marciniak, M., Bijnens, B., Moggridge, J., He, F. J., Antonios, T. F., MacGregor, G. A., and Sutherland, G. R. (2007). Regional left ventricular deformation and geometry analysis provides insights in myocardial remodelling in mild to moderate hypertension. *European Journal of Echocardiography*, 9(4):501–508.

Codreanu, I., Pegg, T. J., Selvanayagam, J. B., Robson, M. D., Rider, O. J., Dasanu, C. A., Jung, B. A., Taggart, D. P., Golding, S. J., Clarke, K., et al. (2014). Normal values of regional and global myocardial wall motion in young and elderly individuals using navigator gated tissue phase mapping. *Age*, 36(1):231–241.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

de Vila, M. H., Attar, R., Pereanez, M., and Frangi, A. F. (2018). Multi-x, a state-of-the-art cloud-based ecosystem for biomedical research. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1726–1733. IEEE.

Deviggiano, A., Carrascosa, P., De Zan, M., Capuñay, C., Deschle, H., and Rodríguez Granillo, G. A. (2016). Wall thickness and patterns of fibrosis in hypertrophic cardiomyopathy assessed by cardiac magnetic resonance imaging. *Revista Argentina de Cardiología*, 84(3).

Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C., and Iyengar, S. (2016). Computational health informatics in the big data age: a survey. *ACM Computing Surveys (CSUR)*, 49(1):12.

Frangi, A. F., Niessen, W. J., and Viergever, M. A. (2001). Three-dimensional modeling for functional analysis of cardiac images, a review. *IEEE transactions on medical imaging*, 20(1):2–5.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5:13.

Heller, G. V., Cerqueira, M. D., Weissman, N. J., Dilsizian, V., Jacobs, A. K., Kaul, S., Laskey, W. K., Pennell, D. J., Rumberger, J. A., Ryan, T., et al. (2002). 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(2):240–245.

Kanza, R. E., Higashino, H., Kido, T., Kurata, A., Saito, M., Sugawara, Y., and Mochizuki, T. (2007). 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(3):119–126.

Klinke, V., Muzzarelli, S., Lauriers, N., Locca, D., Vincenti, G., Monney, P., Lu, C., Nothnagel, D., Pilz, G., Lombardi, M., et al. (2013). 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(1):55.

Lardo, A., Fayad, Z. A., Chronos, N., and Fuster, V. (2004). *Cardiovascular magnetic resonance: established and emerging applications*. Taylor & Francis.

Le Ven, F., Bibeau, K., De Larochellière, É., Tizón-Marcos, H., Deneault-Bissonnette, S., Pibarot, P., Deschepper, C. F., and Larose, É. (2015). 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(9):981–990.

Medrano-Gracia, P., Cowan, B. R., Suinesiaputra, A., and Young, A. A. (2015). Challenges of cardiac image analysis in large-scale population-based studies. *Current cardiology reports*, 17(3):9.

Petersen, S. E., Aung, N., Sanghvi, M. M., Zemrak, F., Fung, K., Paiva, J. M., Francis, J. M., Khanji, M. Y., Lukaschuk, E., Lee, A. M., et al. (2017). 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(1):18.

Petersen, S. E., Matthews, P. M., Bamberg, F., Bluemke, D. A., Francis, J. M., Friedrich, M. G., Leeson, P., Nagel, E., Plein, S., Rademakers, F. E., et al. (2013). Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK biobank-rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):46.

Petersen, S. E., Matthews, P. M., Francis, J. M., Robson, M. D., Zemrak, F., Boubertakh, R., Young, A. A., Hudson, S., Weale, P., Garratt, S., et al. (2015). UK Biobanks cardiovascular magnetic resonance protocol. *Journal of cardiovascular magnetic resonance*, 18(1):8.

Prasad, M., Ramesh, A., Kavanagh, P., Tamarappoo, B. K., Nakazato, R., Gerlach, J., Cheng, V., Thomson, L. E., Berman, D. S., Germano, G., et al. (2010). 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(2):317–327.

Puntmann, V. O., Yap, Y. G., McKenna, W., and Camm, A. J. (2010). Significance of maximal and regional left ventricular wall thickness in association with arrhythmic events in patients with hypertrophic cardiomyopathy. *Circulation Journal*, 74(3):531–537.

Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., Ahmed, M., Aksut, B., Alam, T., Alam, K., et al. (2017). Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American College of Cardiology*, 70(1):1–25.

Tobon-Gomez, C., Sukno, F. M., Butakoff, C., Huguet, M., and Frangi, A. F. (2012). Automatic training and reliability estimation for 3D ASM applied to cardiac MRI segmentation. *Physics in Medicine & Biology*, 57(13):4155.

Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D., and Glocker, B. (2017). Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging*.

Van Assen, H. C., Danilouchkine, M. G., Frangi, A. F., Ordás, S., Westenberg, J. J., Reiber, J. H., and Lelieveldt, B. P. (2006). SPASM: a 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data. *Medical Image Analysis*, 10(2):286–303.

Zhang, L., Gooya, A., Dong, B., Hua, R., Petersen, S. E., Medrano-Gracia, P., and Frangi, A. F. (2016). Automated quality assessment of cardiac MR images using convolutional neural networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 138–145. Springer.

Zheng, Q., Delingette, H., Duchateau, N., and Ayache, N. (2018). 3D consistent & robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE Transactions on Medical Imaging*.