

High Throughput Computation of Reference Ranges of Biventricular Cardiac Function on the UK Biobank Population Cohort

Rahman Attar¹, Marco Pereañez¹, Ali Gooya¹, Xènia Albà², Le Zhang¹,
Stefan K. Piechnik³, Stefan Neubauer³, Steffen E. Petersen⁴,
Alejandro F. Frangi¹

¹Center for Computational Imaging and Simulation Technologies in Biomedicine, Department of Electronic and Electrical Engineering, The University of Sheffield, UK.

²Center for Computational Imaging and Simulation Technologies in Biomedicine, Universitat Pompeu Fabra, Barcelona, Spain.

³Oxford Center for Clinical Magnetic Resonance Research (OCMR), Division of Cardiovascular Medicine, University of Oxford, John Radcliffe Hospital, UK.

⁴Cardiovascular Medicine at the William Harvey Research Institute, Queen Mary University of London and Barts Heart Center, Barts Health NHS Trust, UK.

Abstract. The exploitation of large-scale population data has the potential to improve healthcare by discovering and understanding patterns and trends within this data. To enable high throughput analysis of cardiac imaging data automatically, a pipeline should comprise quality monitoring of the input images, segmentation of the cardiac structures, assessment of the segmentation quality, and parsing of cardiac functional indexes. We present a fully automatic, high throughput image parsing workflow for the analysis of cardiac MR images, and test its performance on the UK Biobank (UKB) cardiac dataset. The proposed pipeline is capable of performing end-to-end image processing including: data organisation, image quality assessment, shape model initialisation, segmentation, segmentation quality assessment, and functional parameter computation; all without any user interaction. To the best of our knowledge, this is the first paper tackling the fully automatic 3D analysis of the UKB population study, providing reference ranges for all key cardiovascular functional indexes, from both left and right ventricles of the heart. We tested our workflow on a reference cohort of 800 healthy subjects for which manual delineations, and reference functional indexes exist. Our results show statistically significant agreement between the manually obtained reference indexes, and those automatically computed using our framework.

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. Thus, mechanisms for early and reliable quantification of cardiac function is of utmost importance.

Analysis and interpretation of cardiac structural and functional indexes in large-scale population image data can reveal patterns and trends across population groups, and allow insights into risk factors before CVDs develop. UKB is one of the world’s largest population-based prospective studies, established to investigate the determinants of disease.

In terms of population sample size, experimental setup, and quality control, the most reliable reference ranges for cardiovascular structure and function in adult caucasians aged 45-74 found in the literature are those reported in [2]. In [2], cardiovascular magnetic resonance (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 are used in this paper to validate the proposed workflow.

In this paper, we present a fully automatic 3D image parsing workflow with quality control modules to analyse CMR images in the UKB and corroborate their validity compared to their manual counterpart. The proposed workflow is capable of segmenting the cardiac ventricles and generating clinical reference ranges that are statistically comparable to those obtained by human observers. 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 large CMR datasets.

2 Methods

Figure 1 shows the architecture of the proposed workflow addressing the issue of large-scale analysis of CMR images. It consists of eight main modules to analyse every single subject of the database. To create a modular workflow and enable processing multiple subjects in parallel, a workflow manager software package is required. This provides an infrastructure for the set-up, performance and monitoring of a defined sequence of tasks, regardless their programming language. In our implementation, the Nipype package [3] has been used. It allows us to combine a heterogeneous set of software packages within a single and highly efficient workflow, processing several subjects in parallel using cloud computing platforms provided by Amazon (high performance processors and S3 storage services).

2.1 Data Organisation (DO)

The Data Organisation (DO) module was developed to hierarchically organise image series from raw DICOM data. It is important to organise the data files to minimise redundancy and inconsistency. As a result, the organised data provides improved searchability and identification of contents. Clear, descriptive, and unique file names have been used to reflect the contents of the file, uniquely

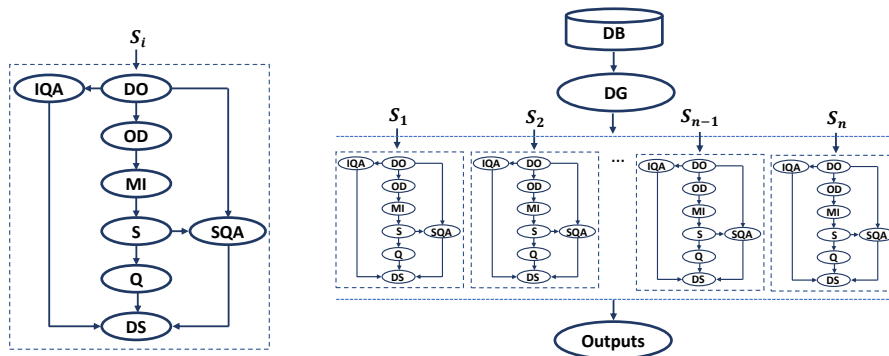


Figure 1. The proposed fully automatic image parsing workflow for the analysis of cardiac ventricles in parallel. Left: The workflow includes the following modules: DO: Data Organisation, IQA: Image Quality Assessment, OD: Organ Detection, MI: Model Initialisation, S: Segmentation, SQA: Segmentation Quality Assessment, Q: Quantification, DS: Data Sink. Right: The quantitative functional analysis of a large database in parallel mode. DB: Database, DG: Data Grabber, n : number of subjects, and S_i : i^{th} subject of the dataset.

identify the data, and enable precise accessibility and data retrieval. Each subject's DICOM data are organised according to cardiac cycle phase, and into short axis (SAX) and long axis (LAX) views.

2.2 Image Quality Assessment (IQA)

Low image quality can not be fully avoided, particularly in large-scale imaging studies. To ensure that the quality of collected data is optimal for statistical analysis, having an IQA module is of paramount. This allows the automatic detection of abnormal images, whose analysis would otherwise impair the 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 has a major impact on the accuracy of quantitative parameters of cardiac function, our IQA module is designed to detect missing apical and basal slices of the CMR input. Thus, every top and bottom short axis view of input image volumes is analysed using two convolutional neural networks, each particularly trained for detection of missing slices in the basal and/or apical positions. The details of the architecture used can be found in [4].

2.3 Organ Detection (OD)

To segment the image, we use a Sparse Active Shape Models framework (SPASM) [5], which requires model initialisation. We achieve this automatically by extending the method proposed in [6] for LV initialisation to biventricular initialisation.

In [6], the location of the LV is determined by a rough estimation of the intersection of slices from different views (SAX and LAX). 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 the final landmark positions. This method is LV specific and therefore we have extended it to take into account image features corresponding to the RV, and obtain optimal initialisations for biventricular segmentation.

2.4 Model Initialization (MI)

The landmarks obtained in Sec. 2.3 are used 1) to suitably place the initial shape inside the image volume (translation), 2) to scale the initial shape along the main axis of the heart (scaling); and 3) to define the initial orientation of the heart based on the relative position of the mitral valve (rotation). These initial pose parameters are estimated by registering the obtained landmarks to their corresponding points on the mean model shape. As we segment all timepoints in the CINE sequence, we initialise the first image timepoint with the model mean, however, subsequent cardiac phases are initialised with the resulting segmentation from the previous timepoint.

2.5 Segmentation (S)

Cardiac LV and RV segmentation is performed with a modified 3D-SPASM segmentation method [5]. The main components of the 3D-SPASM are a Point Distribution Model (PDM), an Intensity Appearance Model (IAM), and a Model Matching Algorithm (MMA).

In this work, the PDM is a surface mesh representing the endocardial and epicardial surfaces for the LV and the endocardial surface for the RV. The PDM is built during training by applying Principal Component Analysis to a set of aligned shapes and maintaining eigenvectors corresponding to a predefined percentage of shape variability. The learned shape variability can be modeled as $\hat{\mathbf{x}} = \bar{\mathbf{x}} + \Phi \mathbf{b}$ where $\hat{\mathbf{x}}$ is a shape model instance, $\bar{\mathbf{x}}$ is the mean shape, Φ is an eigenvector matrix and \mathbf{b} is a vector of scaling values for each principal component. By modifying \mathbf{b} , we can generate shapes from the shape distribution.

The IAM is trained by learning the graylevel intensity distribution along perpendiculars to boundary points on the cardiac shape. An appearance mean and covariance matrix $\{\bar{\mathbf{g}}, \Sigma_{gg}\}$ is computed for each landmark by sampling the intensity around each point over the image training set.

The last element of the segmentation process is the MMA, whose role is iterating between finding the optimal location of boundary points by distance minimisation between sampled image profiles and the IAM, and projection of these points onto the valid shape space defined by the PDM.

2.6 Segmentation Quality Assessment (SQA)

Due to varying image quality, image artefacts, or extreme anatomical variations found in large-scale studies it is essential to have a self-verification capabilities

to automatically detect incorrect results, either to reprocess those images, or disregard them. This becomes even more important when automated segmentation methods are applied to large-scale datasets, and the segmentation results are to be used for further statistical population analysis [7]. In our pipeline we incorporate the SQA proposed in [6]. The SQA uses Random Forest classifiers trained on intensity features associated to blood pool and myocardium, and is able to detect successful segmentations.

2.7 Quantification (Q)

After successful SQA, we compute a thorough set of functional parameters based on blood-pool and myocardial volumes. To reproduce the reference ranges reported in [2], our quantification module performs volume computations using the Simpson’s rule. The principle underlying this method is that total volume can be approximated by the summation of stacks of elliptical disks.

3 Experiments and Results

We use the same dataset exploited in [2], and evaluate the performance of the proposed automatic workflow in two ways: 1) applying common metrics for segmentation accuracy assessment i.e., Dice Similarity Coefficient (DSC), Mean Contour Distance (MCD) and Hausdorff distance (HD), against ground truth values obtained through manual delineation by clinicians. 2) comparing cardiac biventricular function indexes derived from manual and automatic segmentations such as ventricular end diastolic/systolic volumes and myocardial mass. Additionally, quantitative evaluation of human performance i.e., the inter-observer variability, is measured among the manual segmentations of 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 segmentation on the same set of subjects to show how close the performance of the automatic segmentation is to human performance and also the performance of the proposed workflow on a large dataset.

Image volumes at end diastolic and end systolic timepoints of 250 random subjects (500 images in total) were used for training the PDM and IAM. The test dataset contains 800 subjects (not included in training) used for evaluation of the proposed automatic approach. The input images and output segmentation contours were automatically quality controlled to ensure that image volumes included both basal and apical slices, and to verify the automatic segmentation results. After IQA, all 800 images were classified as having full coverage. After SQA, 21 segmentations were deemed suboptimal. Since the aim of the results presented in Sec. 3.1 is the evaluation of segmentation accuracy, all 800 segmentation results (including 21 outliers) were included in the statistics. In contrast, those results presented in Sec. 3.2 are based on 779 good quality segmentations, i.e. excluding those deemed suboptimal by SQA.

3.1 Segmentation Accuracy

Table 1 reports mean and standard deviation for DSC, MCD, and HD comparing between automatic and manual segmentations performed on test sets of 50 and 800 subjects never seen before by the PDM and IAM. The set of 50 subjects is the same set used for the evaluation of inter-observer variability. The set of 800 subjects is the same set used to generate reference ranges in [2].

The reported DSC values show excellent agreement (≥ 0.87) between manual delineations and automatic segmentations. MCD errors are smaller than the in-plane pixel spacing range of 1.8 mm to 2.3 mm found in the UKB. Although HD is larger than the in-plane pixel spacing, it is still within an acceptable range when compared with the distance range seen between different human observers. Table 1 shows that the segmentation accuracy of our method is within error ranges observed between different human raters. This indicates that our workflow performs with human-like reliability, and can fully automatically segment large scale datasets where manual inputs are infeasible.

Table 1. Segmentation accuracy expressed in terms of DSC, MCD and HD comparing the automatic (Auto.), manual (Man.), and observers (O1-O3) segmentations. 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)	Auto. vs Man. (n=50)	Auto. vs Man. (n=800)
LV_{endo}	0.94 ± 0.04	0.92 ± 0.04	0.93 ± 0.04	0.93 ± 0.03	0.93 ± 0.04
LV_{myo}	0.88 ± 0.02	0.87 ± 0.03	0.88 ± 0.02	0.88 ± 0.03	0.87 ± 0.03
RV_{endo}	0.87 ± 0.06	0.88 ± 0.05	0.89 ± 0.05	0.87 ± 0.06	0.89 ± 0.05
(b) MCD (mm)					
	O1 vs O2 (n=50)	O2 vs O3 (n=50)	O3 vs O1 (n=50)	Auto. vs Man. (n=50)	Auto. vs Man. (n=800)
LV_{endo}	1.00 ± 0.25	1.30 ± 0.37	1.21 ± 0.48	1.28 ± 0.39	1.17 ± 0.32
LV_{myo}	1.16 ± 0.34	1.19 ± 0.25	1.21 ± 0.36	1.20 ± 0.34	1.16 ± 0.40
RV_{endo}	2.00 ± 0.79	1.78 ± 0.45	1.87 ± 0.74	1.79 ± 0.80	1.81 ± 0.67
(c) HD (mm)					
	O1 vs O2 (n=50)	O2 vs O3 (n=50)	O3 vs O1 (n=50)	Auto. vs Man. (n=50)	Auto. vs Man. (n=800)
LV_{endo}	2.84 ± 0.70	3.31 ± 0.90	3.25 ± 0.96	3.21 ± 0.97	3.21 ± 0.99
LV_{myo}	3.70 ± 1.16	3.82 ± 1.07	3.76 ± 1.21	3.91 ± 1.20	3.92 ± 1.30
RV_{endo}	7.56 ± 5.51	7.35 ± 2.19	7.14 ± 2.20	7.41 ± 4.11	7.31 ± 3.32

3.2 Estimation of Cardiac Function Indexes

We evaluate the accuracy of cardiac function indexes derived from automatic segmentation versus gold standard reference ranges derived from manual segmentation. We calculate 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 Stroke Volume (RVSV) and RV Ejection-Fraction (RVEF)

from automated segmentation and compare them to measurements from manual segmentation.

Table 2 shows excellent agreement between the mean and standard deviation of ventricular parameters of a healthy population obtained through both automatic and manual segmentations. 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 the manual and automatic methods are from the same continuous distribution in terms of clinical indexes. K-S test on different indexes does not reject the null hypothesis of being from same distribution at the 5% significance level.

Table 2. Cardiac function indexes derived from manual (Man.) vs automatic (Auto.) segmentation on 779 subjects. Values indicate mean \pm standard deviation.

	LVEDV (ml)	LVESV (ml)	LVSV (ml)	LVEF (%)	LVM (g)	RVEDV (ml)	RVESV (ml)	RVSV (ml)	RVEF (%)
Man.	144 \pm 34	59 \pm 18	85 \pm 20	60 \pm 6	86 \pm 24	154 \pm 40	69 \pm 24	85 \pm 20	56 \pm 6
Auto.	146 \pm 31	60 \pm 18	86 \pm 18	60 \pm 7	87 \pm 23	154 \pm 40	71 \pm 26	83 \pm 21	54 \pm 7

Figure 2 shows Bland-Altman (top) and correlation (bottom) plots of ventricular parameters computed using the proposed automatic method and the manual reference on the test dataset. 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. The correlation plots and their correlation coefficient (corr) indicate a strong relationship between the manual and automatic approaches.

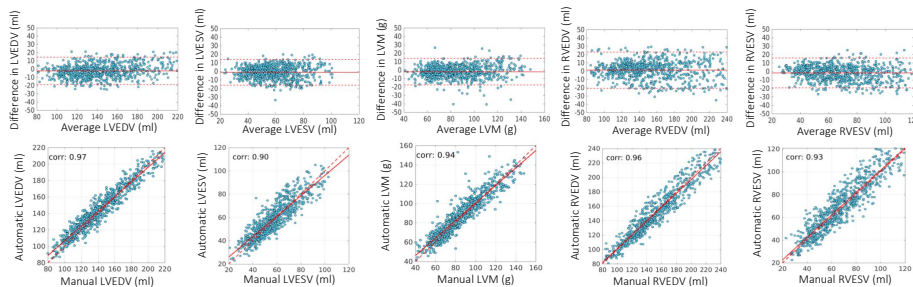


Figure 2. Repeatability of various cardiac functional indexes: manual vs automatic analysis on the test dataset. The first row shows **Bland-Altman** plots. The solid line denotes the mean difference (bias) and the two dashed lines denote ± 1.96 standard deviations from the mean. The second row shows **correlation** plots. The dashed and solid line denote the identity and linear regression lines, respectively.

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 800 healthy subjects for which manual delineations, and reference functional indexes exist. Our results show statistically significant agreement between the manually obtained reference indexes, and those computed automatically using the proposed workflow. As future work, we plan to analyse all available UKB datasets including both healthy and pathological subjects and report the regional and global cardiac function indexes.

Acknowledgements 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.

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*, vol. 70, no. 1, pp. 1–25, 2017.
2. 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*, vol. 19, no. 1, p. 18, 2017.
3. K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh, “Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python,” *Frontiers in neuroinformatics*, vol. 5, p. 13, 2011.
4. L. Zhang, A. Gooya, B. Dong, R. Hua, S. E. Petersen, P. Medrano-Gracia, and A. F. Frangi, “Automated quality assessment of cardiac MR images using convolutional neural networks,” in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 138–145, Springer, 2016.
5. H. C. Van Assen, M. G. Danilouchkine, A. F. Frangi, S. Ordás, J. J. Westenberg, J. H. Reiber, and B. P. Lelieveldt, “SPASM: a 3D-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data,” *Medical Image Analysis*, vol. 10, no. 2, pp. 286–303, 2006.
6. X. Albà, K. Lekadir, M. Pereañez, P. Medrano-Gracia, A. A. Young, and A. F. Frangi, “Automatic initialization and quality control of large-scale cardiac MRI segmentations,” *Medical image analysis*, vol. 43, pp. 129–141, 2018.
7. V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, “Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth,” *IEEE Transactions on Medical Imaging*, 2017.