

# Quality Control-Driven Image Segmentation: Cardiovascular Magnetic Resonance Aortic Cine Imaging

Evan Hann<sup>1</sup>[0000-0002-8911-923X], Luca Biasioli<sup>1</sup>[0000-0002-0452-8756], Qiang Zhang<sup>1</sup>, Iulia A. Popescu<sup>1</sup>, Konrad Werys<sup>1</sup>, Elena Lukaschuk<sup>1</sup>, Valentina Carapella<sup>1</sup>, Jose M. Paiva<sup>2</sup>, Nay Aung<sup>2</sup>, Jennifer J. Rayner<sup>1</sup>, Kenneth Fung<sup>2</sup>, Henrike Puchta<sup>1</sup>, Mihir M. Sanghvi<sup>2</sup>, Niall O. Moon<sup>1</sup>, Katherine E. Thomas<sup>1</sup>, Vanessa M. Ferreira<sup>1</sup>, Steffen E. Petersen<sup>2</sup>, Stefan Neubauer<sup>1</sup>, Stefan K. Piechnik<sup>1</sup>

<sup>1</sup> Oxford Centre for Clinical Magnetic Resonance Research (OCMR), Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom

<sup>2</sup> William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, London, United Kingdom

**Abstract.** Recent progress in fully-automated image segmentation has enabled efficient extraction of clinical parameters in large-scale clinical imaging studies reducing laborious manual processing. However, the current state-of-the-art automatic image segmentation may still fail, especially when it comes to unfamiliar cases. Visual inspection of segmentation quality is often required, thus diminishing the improvements in efficiency. This drives an increasing interest to enhance the overall data processing pipeline with adequate automatic quality scoring. In this work, we present a novel quality control-driven (QCD) framework to provide improved segmentation by utilizing a set of different neural networks. In contrast to the prior segmentation and quality scoring methods, the proposed framework automatically selects the optimal segmentation on-the-fly from the multiple candidate segmentations available, directly utilizing the inherent Dice metrics (DSC) predictions. We train and evaluate the framework on a large-scale cardiovascular magnetic resonance aortic cine image sequences from the UK Biobank Study. The framework achieves segmentation accuracy of mean DSC at 0.966, mean prediction error of DSC within 0.015, and mean error in estimating lumen area less than 17.6 mm<sup>2</sup> for both ascending aorta and proximal descending aorta. The introduced QCD framework can be deployed to integrate the robust extraction of clinical parameters with detection of critical errors in large-scale imaging studies.

**Keywords:** Quality control, Segmentation, Convolutional neural networks.

## 1 Introduction

Aortic distensibility (AoD) is a clinical parameter which measures the bio-elastic function of the aorta. It can serve as an independent predictor for cardiovascular morbidity and mortality [1]. The calculation of AoD involves measurement of the aortic lumen area over a cardiac cycle, from diastole to systole. In the current clinical practice, this requires manual contouring of both ascending aorta (AA) and proximal descending

aorta (PDA) in individual images of the cardiovascular magnetic resonance (CMR) cine sequence.

Manual segmentation is time-consuming, labor-intensive, and subject to inter and intra-observer variability, especially when it comes to large-scale imaging studies such as the UK Biobank (UKBB), which aims to acquire CMR images from 100,000 participants [2]. Hence, these large-scale studies can benefit from using automated image segmentation to alleviate the burden of manual processing.

However, the issue of quality control needs to be addressed for potential deployment of automated segmentation to large-scale imaging studies. The current state-of-the-art segmentation methods can still fail when it comes to difficult cases affected by image quality or pathologies [3]. Hence, it is important to detect any critical inaccuracies, which can potentially lead to misdiagnosis or incorrect research conclusion. The current clinical practice of segmentation quality control requires visual inspection, which diminishes the benefits of efficiency brought forth by automated segmentation. This poses a demand for automation of segmentation quality control integrated in a fully-automated image analysis pipeline, which can efficiently and reliably extract clinical parameters such as AoD in large-scale clinical studies.

### 1.1 Related Works

**Fully-automatic aortic image segmentation methods** have been proposed [4, 5]. A recurrent neural network (RNN) in [4] was trained on 400 scans with label propagation and weighted loss technique to mitigate the sparse annotation problem, as only systolic and diastolic frames were manually annotated in each image sequence. Subsequently, the trained RNN was evaluated in a small-scale dataset of 100 scans. Another approach was proposed in [5] using random forest (RF) localization of the aorta, with a large-scale evaluation. First, potential locations of AA and PDA are detected using Circular Hough Transform (CHT). This is followed by RF classifications based on 18 spatial, intensity, and shape features to select the most probable locations of AA and PDA. This fully-automatic localization method can initialize semi-automatic segmentation methods. It was tested on 3900 scans in the UKBB imaging study to achieve detection accuracy over 99% for both AA and PDA. However, neither of the publications included a quality control mechanism to predict the accuracy of segmentations.

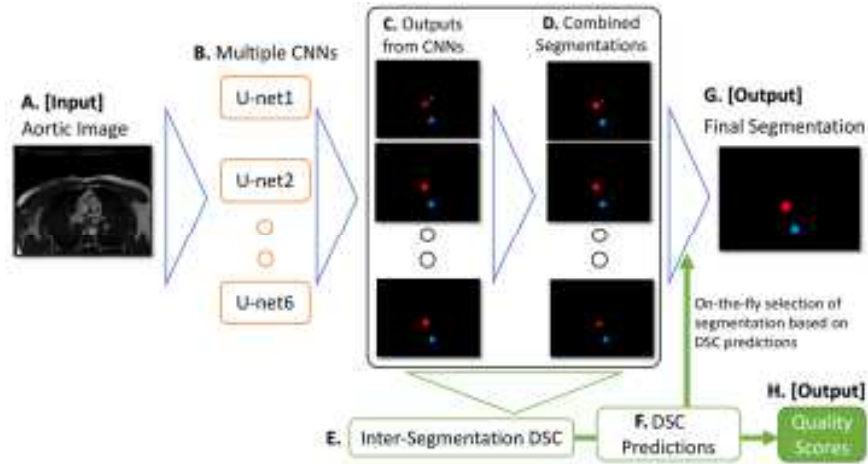
**Automatic Dice metrics predictions** have been proposed to address the segmentation quality control in the absence of manual segmentation. Kohlberger et al. [6] proposed an automated quality scoring of segmentation using machine learning with 42 hand-crafted features evaluated against Dice metrics (DSC). More recently, a framework based on Reverse Classification Accuracy (RCA) [7, 8] has been proposed to predict DSC and other metrics for CMR segmentation. The RCA framework requires registration of the input image and the corresponding segmentation to a database of reference images, with available ground truth segmentations. Robinson et al. [9] proposed a simple CNN-based method trained to predict the DSC of segmentations generated by RF-based algorithms. Another CNN-based framework [10] was proposed to predict segmentation DSC using Monte Carlo sampling. With the use of random dropout unit at

test time, the CNN generates several different segmentations for the same input to predict segmentation quality. In these prior works, DSC predictions have not been used to optimize segmentation performance.

## 1.2 Contributions

In this work, we present a novel quality control-driven (QCD) image analysis framework, which utilizes multiple neural networks to integrate segmentation with inherent quality scoring. The concept of the QCD framework hinges on automated selection of the best final segmentation from multiple candidate models based on accurate DSC prediction, rather than only passive reporting as in [6–10]. We evaluate the effectiveness of QCD on a large-scale dataset of aortic cine image sequences from the UKBB imaging study.

## 2 Methods and Material



**Fig. 1.** The overview of quality control-driven (QCD) framework, which feeds the same (A) aortic image frame to (B) multiple convolutional neural networks (U-Nets). (C) The multiple segmentations generated by the U-Nets are summed up and thresholded to form (D) additional combined segmentations. (E) The inter-segmentation DSC matrix is calculated between all segmentation candidates. The matrix is fed into previously established regression model in order to obtain (F) individual DSC prediction for each candidate. (G) The final segmentation is selected on-the-fly for its highest predicted DSC, which is also provided as (H) the final quality scores of the segmentation.

### 2.1 Candidate Segmentation Models

**Multiple Convolutional Neural Networks:** U-Nets [11], with different depths, are implemented to perform image segmentations of AA and PDA. In this work we use 6

U-Nets with number of skip connections from 1 to 6 (U-Net 1 to U-Net 6 in Fig 1B). Such differences in the hyperparameters are intended to introduce variation in segmentation performance, which is exploited for segmentation quality control.

**Combined Segmentations:** Statistical rank filters are used to combine multiple U-Net segmentations to generate additional segmentations for improved robustness at small additional computation cost. In contrast to a typical rank filter which processes a single image, the rank filters used in this work are applied in a pixel-wise fashion across all 6 U-Net segmentations, such that

$$CS_t(u, v) = \begin{cases} 1, & \sum_{net \in Nets} S_{net}(u, v) \geq t \\ 0, & otherwise. \end{cases} \quad (1)$$

where  $CS_t$  is a combined segmentation with thresholding parameter  $t$ ,  $S_{net}$  is the segmentation output by a U-Net  $net$ , and  $(u, v)$  is a pixel in the segmentation. Hence, for each input of aortic image, there are in total 12 candidate segmentations including U-Nets and combined segmentations for each aorta.

## 2.2 Quality Scoring and Quality Control-Driven Segmentation

**Automatic Quality Scoring:** Availability of multiple segmentations (Fig. 1 C and D) allows comparison between candidates. We hypothesize that the candidates vary more in segmentation when it comes to difficult cases affected by image quality or pathologies. DSC scores are calculated between all pairs of candidate segmentations. These inter-segmentation DSCs are used to predict the target DSC, which is defined to be the DSC compared with the ground truth manual segmentation. The DSC prediction is performed through a linear regression model, established in the training data to map inter-segmentation DSCs to the target DSC for each segmentation model.

**Quality Control-Driven Segmentation:** The QCD framework utilizes the DSC prediction to select the final segmentation. For each aorta in an aortic image frame, 12 candidate segmentations are generated. Each of these candidates is assigned a predicted DSC through the automatic quality scoring. Then, the framework selects on-the-fly the final segmentation with the highest predicted DSC among all candidates. This is to further improve the overall accuracy and robustness of segmentation.

## 2.3 Data and Annotations

5028 CMR aortic cine image sequences acquired in the UKBB are available for training and testing of the QCD framework. In each image sequence, 100 frames across a cardiac cycle were acquired, with pixel dimension of  $240 \times 196$  and resolution of  $1.53 \times 1.53 \text{ mm}^2$ .

The manually-validated segmentations of AA and PDA in the aortic image sequences were generated prior to this work in a semi-automatic fashion by 13 image analysts, using both random forest (RF) localization implemented [5] and 2D active contour [12] (RF-AC). The RF method selected most probable AA and PDA locations to initialize the active contour models. Segmentations generated by the active contours were then visually validated and manually corrected by the image analysts.

Due to the large volume of the dataset (502,800 image frames in total), only frames at systole and diastole (~15 out of 100 frames) were manually validated and corrected to reduce the workload on the image analysts. This leads to a sparse annotation problem similarly reported in [4]. To mitigate this problem, all generated segmentations are used to train the QCD framework, but only manually-validated segmentations are used for evaluation.

## 2.4 Evaluation

The objectives of the evaluation are 4-fold: (1) To compare the QCD segmentation with all available candidate segmentations; (2) To compare the QCD segmentation with the RF-AC segmentation, which generated part of the training annotations; (3) To validate the quality scoring on all segmentation models, with varying performance; (4) To evaluate the performance of the QCD framework in a large-scale testing dataset, 10 times larger than the training dataset. The evaluation is performed on the validation dataset for objectives 1-3, and the testing dataset for objective 4, respectively consisting of 400 and 4228 image sequences.

Segmentation performance is evaluated using Dice metrics (DSC). Mean absolute error (MAE) and Pearson correlation ( $r$ ) between the ground truth DSC and Predicted DSC are calculated to evaluate the quality scoring. Agreement in Aortic lumen area (number of pixels in segmentation multiplied by image resolution) estimated with automated and manual annotations is evaluated in terms of MAE.

## 3 Experiments and Results

### 3.1 Implementation

The framework was implemented in Python, with TensorFlow. Similar to [4], 400 image sequences were used to train the framework. Each of the 6 U-Nets was independently trained in a batch size of 50 frames for 201,200 iterations. The training took 71 hours in total on a desktop computer with a Nvidia Titan X GPU. On average, the framework took 67 seconds to segment and quality score cine of 100 frames (0.67s/image).

### 3.2 Comparison of QCD Segmentation to Baseline Models

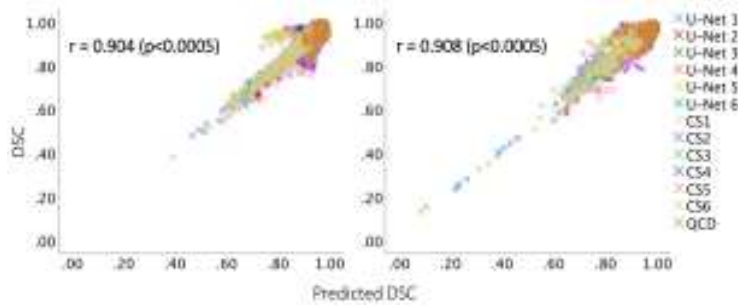
All segmentation models were evaluated for their DSC performance in the validation data (Table 1). RF-AC and QCD respectively achieved the highest DSC for the AA and PDA segmentation. However, RF-AC achieved a relatively low DSC (0.950) for PDA. Table 1 also shows that QCD and RF-AC having 99.7% of the segmentations achieving DSC over 0.9. Among all models, QCD had the best overall segmentation performance in the validation data.

To further evaluate QCD and RF-AC, the minimum DSC was obtained for each model. The QCD framework obtained minimum DSC at 0.834 for AA and 0.817 for

PDA, outperforming RF-AC with DSC 0.000 for AA and 0.413 for PDA. Thus, QCD showed performance gain in segmentation robustness over RF-AC.

**Table 1.** Mean DSC between manual and automatic segmentation, with percentages of segmentations achieving DSC over 0.9, for each model evaluated in the validation data

Model	Mean DSC		Percentage of DSC > 0.9 (%)	
	AA	PDA	AA	PDA
U-Net 1	0.918	0.926	77.4	83.7
U-Net 2	0.949	0.957	97.5	98.9
U-Net 3	0.954	0.961	99.4	99.3
U-Net 4	0.951	0.955	98.8	98.7
U-Net 5	0.953	0.955	99.4	98.5
U-Net 6	0.953	0.956	99.5	99.0
CS1	0.937	0.942	93.7	92.5
CS2	0.964	0.964	98.8	99.2
CS3	0.967	<b>0.966</b>	99.6	99.6
CS4	0.966	<b>0.966</b>	99.6	99.6
CS5	0.958	0.962	99.3	99.4
CS6	0.924	0.934	85.8	90.3
<b>QCD</b>	0.967	<b>0.966</b>	<b>99.7</b>	<b>99.7</b>
<b>RF-AC</b>	<b>0.968</b>	0.950	<b>99.7</b>	<b>99.7</b>



**Fig. 2.** Scatter plots of predicted DSC (x-axis) and DSC (y-axis) for AA (left) and PDA (right) in the validation data, with correlation coefficients ( $r$ ), and  $p$ -values reported. Low DSC scores of poor segmentations output by U-Net 1 and CS6 were accurately predicted.

### 3.3 Quality Scoring of Segmentations

The segmentation quality scoring was evaluated with all candidate segmentations in the validation data. The results show high agreement between DSC and predicted DSC for

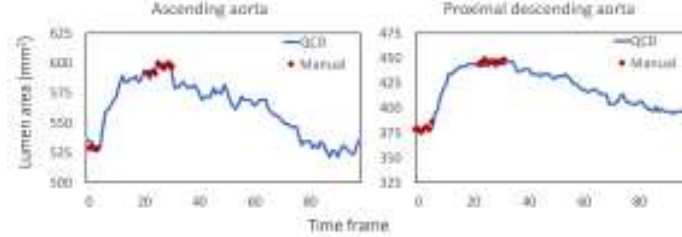
both AA and PDA segmentation, with MAE at 0.009 for AA and 0.012 for PDA, and Pearson correlation over 0.9. The scatter plots (Fig. 2) show that DSC and predicted DSC meet along the identity lines, indicating accurate DSC predictions for segmentations of varying quality.

### 3.4 Large-Scale Testing

Tested on 4228 image sequences, the QCD framework was shown to perform as consistently in the large-scale dataset as in the smaller validation dataset. The segmentation performance, with mean DSC at 0.966 for both AA and PDA (Table 2) was comparable to the validation results. The lumen area estimation was in high agreement with the manual annotations with MAE less than 17.6 mm<sup>2</sup>. Two examples of lumen area curves are shown in Fig. 3. Both curves show consistent lumen area estimation with manual annotations at systole and diastole. In addition, Fig. 4 shows an example in the testing data to demonstrate how differences in candidate segmentations affect the DSC prediction in the QCD framework.

**Table 2.** Evaluation results of QCD framework in the test dataset of 4228 image sequences

Label	Mean DSC	MAE in DSC Prediction	MAE in Lumen Area (mm <sup>2</sup> )
AA	0.966	0.011	17.6
PDA	0.966	0.015	10.5

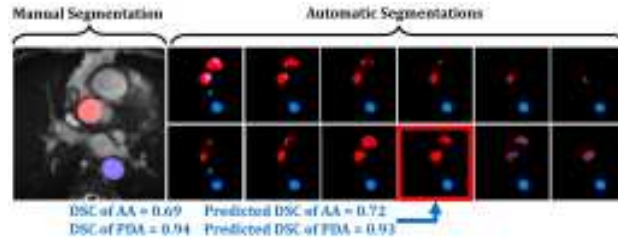


**Fig. 3.** Lumen area curves for AA (left) and PDA (right) estimated by QCD (blue), compared with manually validated ground truth (red).

## 4 Conclusions

In this paper, we presented a novel quality control-driven segmentation framework comprising of different neural networks. In the absence of manual annotations, the framework exploits differences among candidate segmentations in order to predict Dice metrics (DSC), which can be further utilized to select an optimal final segmentation on a per-case basis. Evaluated on a large-scale dataset of aortic cine images, the framework achieved high accuracy in segmentation, quality scoring, and lumen area estimation. This paves the way for fully-automated image analysis pipeline for reliable extraction

of clinical parameters in large-scale clinical studies. Future work will cover a wider range of applications in multiple organs and imaging modalities.



**Fig. 4.** Example of manual segmentation (large panel on the left) with multiple automatic candidate segmentations of AA (red masks) and PDA (blue masks). The panel outlined in a red box shows the selected final segmentation. Its predicted DSC of AA segmentation is low (0.72) due to apparent differences among candidate segmentations, as the AA is affected by suboptimal image quality. In contrast, with the PDA less affected by the image quality, the predicted DSC for PDA is higher (0.93) since there is a higher agreement among candidate models.

## References

1. Redheuil, A. et al.: Proximal aortic distensibility is an independent predictor of all-cause mortality and incident CV events: the MESA study. *JACC*, vol. 64, pp. 2619-2629, 12 2014.
2. Petersen, S.E. et al.: Imaging in population science: Cardiovascular magnetic resonance in 100,000 participants of UK Biobank - Rationale, challenges and approaches. *JCMR*, vol. 15, p. 46, 2013.
3. Bernard, O. et al.: Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved?, *IEEE Transactions on Medical Imaging*, p. 1-1, 2018.
4. Bai, W. et al.: Recurrent Neural Networks for Aortic Image Sequence Segmentation with Sparse Annotations. In *MICCAI*, p. 586-594, 2018.
5. Biasioli L. et al.: Automated localization and quality control of the aorta in cine CMR can significantly accelerate processing of the UK Biobank population data. *PloS one*, vol. 14, p. e0212272, 2019.
6. Kohlberger T. et al.: Evaluating segmentation error without ground truth. In *MICCAI*, p. 528-538 2012.
7. Robinson, R. et al.: Automatic Quality Control of Cardiac MRI Segmentation in Large-Scale Population Imaging. In *MICCAI*, p. 720-727 2017.
8. Robinson R. et al.: Automated Quality Control in Image Segmentation: Application to the UK Biobank Cardiac MR Imaging Study. *JCMR*, vol. 21, pp. 18, 12 2019.
9. Robinson, R. et al.: Subject-level Prediction of Segmentation Failure using Real-Time Convolutional Neural Nets. In *MIDL*, 2018.
10. Roy, A.G. et al.: Inherent brain segmentation quality control from fully convnet monte carlo sampling. In *MICCAI*, pp. 664-672, 2018.
11. Ronneberger, O. et al.: U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234-241, 2015.
12. Kass, M. et al.: Snakes: Active contour models. In *IJCV*, vol. 1, pp. 321-331, 1 1988.