Quality Control-Driven Image Segmentation Towards Reliable Automatic Image Analysis in Large-Scale Cardiovascular Magnetic Resonance Aortic Cine Imaging

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

[1] Oxford Centre for Clinical Magnetic Resonance Research (OCMR), Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom
[2] 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 atypical cases. Visual inspection of segmentation quality is often required, thus diminishing the improvements in efficiency. This drives an increasing need to enhance the overall data processing pipeline with robust automatic quality scoring, especially for clinical applications. We present a novel quality control-driven (QCD) framework to provide reliable segmentation using 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 similarity coefficient (DSC) predictions. We trained and evaluated the framework on a large-scale cardiovascular magnetic resonance aortic cine image sequences from the UK Biobank Study. The framework achieved segmentation accuracy of mean DSC at 0.966, mean prediction error of DSC within 0.015, and mean error in estimating lumen area $\leq$ 17.6 mm^2 for both ascending aorta and proximal descending aorta. This novel QCD framework successfully integrates the automatic image segmentation along with detection of critical errors on a per-case basis, paving the way towards reliable fully-automatic extraction of clinical parameters for 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]. In current clinical practice, this requires cardiovascular magnetic resonance (CMR) trans-axial cine images at the level of the pulmonary artery, with manual contouring of the cross-sectional lumen area of the ascending aorta (AA) and the proximal descending aorta (PDA) over a cardiac cycle, from diastole to systole.

Manual segmentation is time-consuming, labor-intensive, and subject to inter and intra-observer variability, especially in large-scale imaging studies, such as the UK Biobank (UKBB), aiming to acquire CMR images from 100,000 participants [2]. Large-scale studies can benefit from automated image segmentation, achieve not only efficient image segmentation, but also improved consistency and objectivity for diagnosis.

However, the issue of quality control needs to be addressed before deployment of automated segmentation to large-scale imaging studies and clinical applications. The current state-of-the-art segmentation methods can still fail [3], especially in cases affected by poor image quality or pathologies. It is important to detect any critical inaccuracies, which can potentially lead to misdiagnosis or incorrect research conclusion. 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 automated quality control to be integrated in fully-automated image analysis pipelines, to efficiently and reliably extract clinical parameters.

1.1 Related Works

Fully-automatic aortic image segmentation methods without quality control 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 (3900 image sequences) evaluation. First, potential locations of AA and PDA were detected using Circular Hough Transform (CHT), 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 in the UKBB imaging study to achieve detection accuracy over 99% for both AA and PDA. However, neither approach 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 handcrafted features evaluated against Dice similarity coefficient (DSC). More recently, a framework based on Reverse Classification Accuracy (RCA) [7, 8] was proposed to predict DSC and other metrics for CMR image 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. However, 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 and quality scoring on a per-case basis. The QCD framework automatically selects of the best final segmentation from multiple candidate models based on accurate DSC predictions, 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

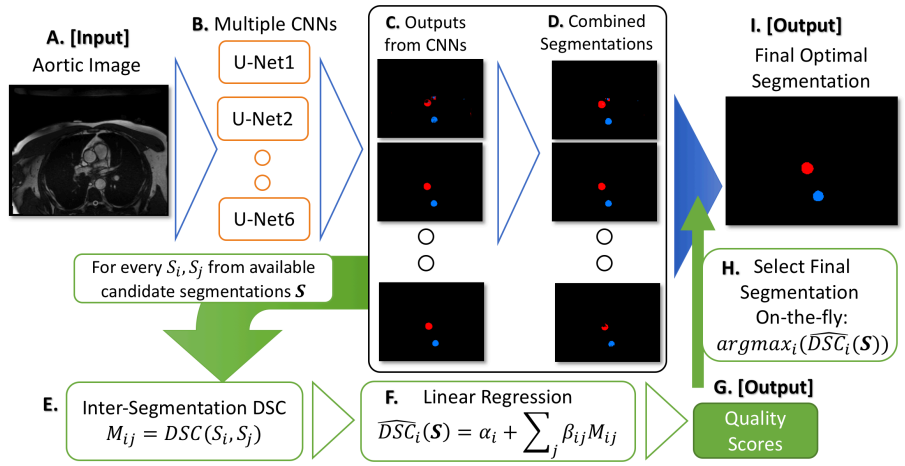

Fig. 1. The overview of the quality control-driven (QCD) framework, which feeds the same aortic CMR image frame (A) to multiple convolutional neural networks (U-Net 1-6) (B). Multiple segmentations (C) generated by the U-Nets are summed up and thresholded to form additional combined segmentations (D). The inter-segmentation DSC matrix (E) is calculated among all segmentation candidates, and fed into a previously established regression model (F) to obtain individual DSC prediction (G) for each candidate. The segmentation with the highest predicted DSC (H) among the candidates is selected on-the-fly as the final segmentation (I).

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 (Fig. 1D) 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 $t \in \{1, 2, \dots, 6\}$, 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 section.

2.2 Quality Scoring and Quality Control-Driven Segmentation

Automatic Quality Scoring predicts $DSC(\cdot, S_{GT})$ by comparing multiple candidate segmentations $\boldsymbol{S}$ (Fig. 1C and D) in the absence of the manual segmentation S_{GT}. For each segmentation S_i, DSCs with other candidates S_j of the same input are calculated to form the inter-segmentation DSC matrix $M_{ij} = DSC(S_i, S_j)$ (Fig. 1E), and then used to predict $DSC(S_i, S_{GT})$ through multiple linear regression $\widehat{DSC_i}(\boldsymbol{S}) = \alpha_i + \sum_j \beta_{ij} M_{ij}$ (Fig. 1F), where regression parameters α_i and β_{ij} are optimized for each segmentation model i using the training data. The prediction exploits differences among candidate segmentations, which tend to diverge in more difficult cases (e.g. affected by poor image quality), for which lower predicted DSCs are anticipated. In contrast, a higher predicted DSC is expected when there is higher agreement among candidates.

Quality Control-Driven Segmentation uses the DSC prediction to select the best final segmentation (Fig. 1I). For each aorta section 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 the final segmentation with the highest predicted DSC (Fig. 1H) from all candidates $\boldsymbol{S}$ on-the-fly: $argmax_i(\widehat{DSC_i}(\boldsymbol{S}))$. This is to further improve accuracy and robustness of segmentation by choosing the predicted best on a per-case basis.

2.3 Data and Annotations

The dataset comprises of 5028 CMR aortic cine image sequences acquired in the UKBB. In each image sequence, 100 frames across a cardiac cycle were acquired, with pixel dimension of 240 × 196 and resolution of 1.58 × 1.58 mm^2.

The manually-validated segmentations of AA and PDA were generated prior to this work in a semi-automatic fashion using both random forest (RF) localization [5] and 2D active contour [12]. The RF method selected the 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 13 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 presents a sparse annotation problem, similar to that 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 3-fold: (1) to evaluate the segmentation accuracy of all segmentation models, including the QCD segmentation, using Dice metrics (DSC); (2) to evaluate the accuracy of quality scoring on all candidate segmentations, with varying quality, using mean absolute error (MAE) and Pearson correlation (r) between the ground truth DSC and the predicted DSC; (3) to evaluate the accuracy of segmentation, quality scoring, and clinical parameter estimation using a large-scale testing dataset, 10 times larger than the training dataset. Agreement in aortic lumen area (number of pixels in segmentation scaled by pixel spacing) estimated with automated and manual annotations is evaluated in terms of MAE. The evaluation is performed in the validation dataset (400 image sequences) for objectives 1 and 2, and the testing dataset (4228 image sequences) for objective 3.

3 Experiments and Results

3.1 Implementation

The framework was implemented in Python, with TensorFlow. Similar to [4], 400 CMR 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 cine frames.

3.2 Performance of Segmentation Models

All segmentation models were evaluated for DSC performance in the validation dataset (Table 1). QCD achieved the highest DSC for AA (0.967) and PDA (0.966) segmentation. Similar segmentation accuracy was also achieved by CS3, which was selected by QCD as the best candidate over 60% of the cases. In addition, CS2-5 obtained higher DSCs than any individual U-Nets, showing the benefit of combing multiple neural networks. Moreover, the results (Table 1) showed that QCD obtained the highest percentages ($\geq$ 99.7%) of segmentations achieving DSC over 0.9, offering additional robustness by selecting the best candidate segmentation on a per-case basis. QCD had the best overall segmentation performance in the validation data.

3.3 Quality Scoring of Segmentations

The segmentation quality scoring was evaluated for all candidate segmentations in the validation dataset. The results showed high agreement between DSC and predicted DSC for both AA and PDA segmentation, with MAE of 0.009 for AA and 0.012 for PDA, and Pearson correlations of over 0.9 for both AA and PDA. The scatter plots (Fig. 2) showed that DSC and predicted DSC met along the identity lines, indicating accurate DSC predictions for segmentations of varying quality.

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

	Mean DSC		Percentage of DSC > 0.9	
Model	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**	**0.966**	99.6	99.6
CS4	0.966	**0.966**	99.6	99.6
CS5	0.958	0.962	99.3	99.4
CS6	0.924	0.934	85.8	90.3
QCD	**0.967**	**0.966**	**99.9**	**99.7**

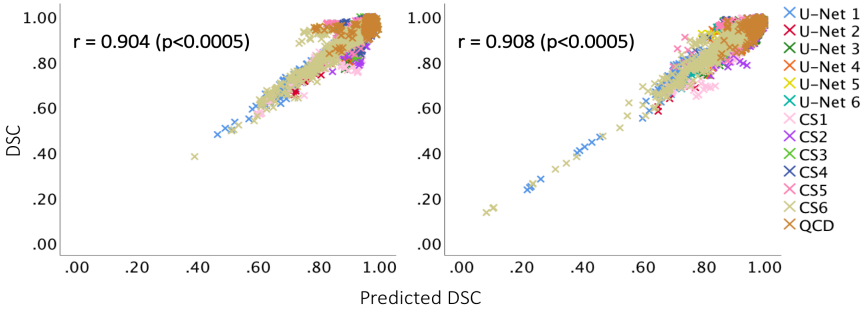

Fig. 2. Scatter plots of predicted DSC (x-axis) and DSC (y-axis) for AA (left) and PDA (right) in the validation data, with correlation coefficients (r), and p-values for all data points reported. Overall good DSC prediction for all candidate segmentations, with varying quality. Low DSC scores of poor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

3.4 Large-Scale Testing

The QCD framework was tested on 4228 image sequences and performed as consistently in the large-scale dataset as in the smaller validation dataset. The segmentation performance, with mean DSC of 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^2 for both AA and PDA (Table 2). 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 influence the DSC predictions 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^2)
AA	0.966	0.011	17.6
PDA	0.966	0.015	10.5

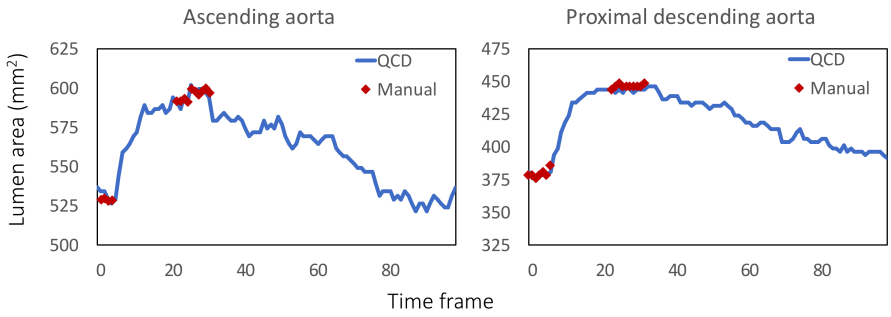

Fig. 3. Lumen area curves for AA (left) and PDA (right) estimated by QCD (blue), compared with manually validated ground truth (red; only in end-diastolic and end-systolic frames).

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 to predict Dice metrics (DSC), which are exploited to select the optimal final segmentation on a per-case basis on-the-fly. 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 for large-scale clinical studies. Future work will cover a wider range of applications in multiple organs and imaging modalities.

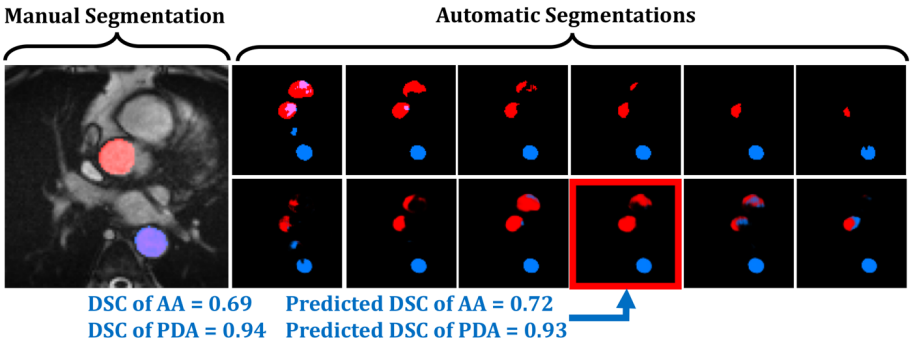

Fig. 4. Example of a poorly-planned aortic cine image (too far below the main pulmonary artery). Manual segmentation (left large panel), with multiple automatic candidate segmentations of AA (red masks) and PDA (blue masks) are shown. For the final selected segmentation (outlined in red), the predicted DSC of AA segmentation is low (0.72) due to apparent differences among candidate segmentations, as AA was affected by poor image quality; most of the automatic segmentation includes parts of the right ventricle. In contrast, PDA was less affected; the predicted DSC was higher (0.93), as there was higher agreement among candidate models.

Acknowledgements This study was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre at The Oxford University Hospitals, University of Oxford, UK. Authors acknowledge support from the British Heart Foundation Centre of Research Excellence, and donation of GPU from NVIDIA Corp.

References

1. Redheuil, A. et al.: Proximal aortic distensibility is an independent predictor of all-cause mortality and incident CV events: the MESA study. J. Am. Coll. Cardiol. 64, 2619–2629 (2014). https://doi.org/10.1016/j.jacc.2014.09.060.
2. Petersen, S.E. et al.: Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - Rationale, challenges and approaches. Journal of Cardiovascular Magnetic Resonance. 50, 46 (2013). https://doi.org/10.1186/1532-429X-15-46.
3. Bernard, O. et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging. 37, 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502.
4. Bai, W. et al.: Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In: MICCAI. pp. 586–594 (2018). https://doi.org/10.1007/978-3-030-00937-3_67.
5. Biasiolli, 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. 14, e0212272 (2019). https://doi.org/10.1371/journal.pone.0212272.
6. Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L.: Evaluating segmentation error without ground truth. In: MICCAI. pp. 528–36 (2012). https://doi.org/10.1007/978-3-642-33415-3_65.
7. Robinson, R. et al.: Automatic quality control of cardiac MRI segmentation in large-scale population imaging. In: MICCAI. pp. 720–727 (2017). https://doi.org/10.1007/978-3-319-66182-7_82.

8. Robinson, R. et al.: Automated quality control in image segmentation: application to the UK biobank cardiac MR imaging study. J. Cardiovasc. Magn. Reson. 21, 18 (2019). https://doi.org/10.1186/s12968-019-0523-x.
9. Robinson, R. et al.: Subject-level prediction of segmentation failure using real-time convolutional neural nets. In: MIDL. pp. 3–5 (2018).
10. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C.: Inherent brain segmentation quality control from fully convnet monte carlo sampling. In: MICCAI. pp. 664–672 (2018). https://doi.org/10.1007/978-3-030-00928-1_75.
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28.
12. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. Int. J. Comput. Vis. 1, 321–331 (1988). https://doi.org/10.1007/BF00133570