Deep Ensemble Learning-Based Quality Control for Automatic Segmentation in Cardiovascular Magnetic Resonance Imaging

Evan Hann

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2020

Acknowledgement

First of all, I would like to express my gratitude and appreciation to all my supervisors (Prof. Stefan Piechnik, Prof. Vanessa Ferreira, and Prof. Stefan Neubauer), for giving me the precious opportunity and the generous guidance to read a DPhil degree in the University of Oxford Centre for Clinical Magnetic Resonance Research (OCMR).

During my DPhil research, I have learned and received support from many colleagues in OCMR. I would like to thank all of you, especially the current and the former members of the Image Analysis Group: Mrs. Elena Lukaschuk, Ms. Henrike Puchta, Dr. Iulia Popescu, Dr. Qiang Zhang, Dr. Mayooran Shanmuganathan, Dr. Matthew Burrage, Dr. Luca Biasiolli, Dr. Konrad Werys, Dr. Valentina Carapella, and Dr. Ahmet Barutçu.

The UK Biobank (UKBB) Imaging Component provided material which enabled the development and the evaluation of my research for this thesis. I would like to thank Prof. Steffen Petersen and colleagues from the Queen Mary University of London for the special opportunity to collaborate on the UKBB projects.

Last but not the least, I am deeply grateful to my family and friends, in Hong Kong, Oxford, Cambridge, and London, for their constant support and encouragement, which empowered me to go through the ups and downs of this journey.

For the pursuit of this degree, I acknowledge the generous funding support from the Clarendon Scholarship and the Radcliffe Department of Medicine Scholarship, University of Oxford.

Declaration

The work in this thesis is my own, unless specifically acknowledged.

Prof. Stefan Piechnik, Prof. Vanessa Ferreira, Dr. Iulia Popescu, Dr. Qiang Zhang, and Dr. Konrad Werys provided support in the conceptualisation and the patenting of the proposed quality control-driven framework. For Chapter 2, the annotated data were curated by co-authors in [26], led by Dr. Luca Biasiolli, for the UK Biobank (UKBB) Imaging Component. For Chapter 3, the annotated T1 maps were curated internally during multiple past research studies (cited in Chapter 3) in the Oxford Centre for Clinical Magnetic Resonance Research. Dr. Ahmet Barutçu provided ground truth manual annotations of T1 map image quality. For Chapter 4, Mrs. Elena Lukaschuk provided ground truth manual contours and validated image quality for the material available under the UKBB technical development agreement (PI: Prof. Stefan Piechnik). For Chapter 5, Mrs. Elena Lukaschuk also manually validated image quality of the T1 maps. Prof. Stefan Piechnik provided visual quality assessment scores for the automatic T1-map segmentations.

Abstract

Cardiovascular magnetic resonance (CMR) imaging is a powerful tool for research and clinical applications. To extract useful clinical information from the acquired CMR images, time-consuming and laborious manual delineation of cardiovascular structures is currently required. Despite promising overall performance across medical imaging applications, the current state-of-the-art automated image segmentation methods still fail in some cases, potentially jeopardising the reliability of clinical diagnosis. Thus, it is important to develop not only automation of image segmentation but also quality control of segmentation, to empower efficient and reliable CMR image data analysis.

To address both segmentation and quality control problems, I have developed a novel quality control-driven (QCD) framework in this thesis. Extending upon deep ensemble learning, the framework utilises multiple convolutional neural network-based models to generate segmentation candidates, the agreement of which is exploited via additional regression models to predict segmentation quality measured by Dice similarity coefficient (DSC). The DSC prediction not only provides a quality estimate but also enables a novel approach to select a final, most optimal segmentation on-the-fly from multiple candidates, improving segmentation robustness. Following the DSC prediction, a segmentation quality classification scheme is implemented to alert human operators only when manual intervention is recommended, intended for more efficient allocation of time and labour resources for large-scale image processing pipelines.

Through both quantitative and qualitative evaluation, the QCD framework has demonstrated excellent performance in both segmentation and quality control. More importantly, the framework has been successfully applied across CMR imaging tech-

niques, anatomical structures, and large-scale datasets acquired at different sites, with high adaptability and generalisability. The QCD framework could pave the way towards large-scale automated imaging data analysis pipelines, with both efficiency and reliability, in real-world clinical applications.

List of Publications

Full Papers

1. **Hann, E.**, Popescu, I.A., Zhang, Q., Gonzales, R.A., Barutçu, A., Neubauer, S., Ferreira, V.M., Piechnik, S.K., 2021. Deep Neural Network Ensemble for On-the-Fly Quality Control-Driven Segmentation of Cardiac MRI T1 Mapping. Medical Image Analysis. 102029. https://doi.org/10.1016/j.media.2021.102029 **(Thesis Chapter 3)**

2. **Hann, E.**, Biasiolli, L., Zhang, Q., Popescu, I.A., Werys, K., Lukaschuk, E., Carapella, V., Paiva, J.M., Aung, N., Rayner, J.J., Fung, K., Puchta, H., Sanghvi, M.M., Moon, N.O., Thomas, K.E., Ferreira, V.M., Petersen, S.E., Neubauer, S., Piechnik, S.K.. 2019. Quality Control-Driven Image Segmentation Towards Reliable Automatic Image Analysis in Large-Scale Cardiovascular Magnetic Resonance Aortic Cine Imaging, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer International Publishing, Cham, pp. 750–758. `https://doi.org/10.1007/978-3-030-32245-8_83` **(Thesis Chapter 2)**

3. Biasiolli, L., **Hann, E.**, Lukaschuk, E., Carapella, V., Paiva, J.M., Aung, N., Rayner, J.J., Werys, K., Fung, K., Puchta, H., Sanghvi, M.M., Moon, N.O., Thomson, R.J., Thomas, K.E., Robson, M.D., Grau, V., Petersen, S.E., Neubauer, S., Piechnik, S.K.. 2019. 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. https://doi.org/10.1371/journal.pone.0212272 **(Thesis Chapter 2)**

4. Zhang, Q., **Hann, E.**, Werys, K., Wu, C., Popescu, I., Lukaschuk, E., Barutcu, A., Ferreira, V.M., Piechnik, S.K., 2020. Deep learning with attention supervision for automated motion artefact detection in quality control of cardiac T1-mapping. Artif. Intell. Med. 101955. https://doi.org/10.1016/j.artmed.2020.101955

5. Popescu, I.A., Werys, K., Zhang, Q., Puchta, H., **Hann, E.**, Lukaschuk, E., Ferreira, V.M., Piechnik, S.K., 2021. Standardization of T1-mapping in cardiovascular magnetic resonance using clustered structuring for benchmarking normal ranges. Int. J. Cardiol. 326, 220–225. https://doi.org/10.1016/j.ijcard.2020.10.041

6. Werys, K., Dragonu, I., Zhang, Q., Popescu, I., **Hann, E.**, Puchta, H., Kubik, A., Polat, D., Wu, C., Moon, N.O., Barutcu, A., Ferreira, V.M., Piechnik, S.K.. 2020. Total Mapping Toolbox (TOMATO): An open source library for cardiac magnetic resonance parametric mapping. SoftwareX 11, 100369. https://doi.org/10.1016/j.softx.2019.100369

Patents

1. **Hann, E.**, Piechnik, S.K., Popescu, I.A., Zhang, Q., Werys, K., Ferreira, V.M. "Method and Apparatus for Quality Prediction". Patent application submitted to Oxford University Innovation (project 16045) PCT/GB2020/050249 , Filed 4 February 2020 **(Thesis Chapter 2 and 3)**

2. Zhang, Q., Piechnik, S.K., Ferreira, V.M., **Hann, E.**, Popescu, I.A. "Method and apparatus for enhancing medical images", UK Patent Application No. GB1912701.8, Oxford University Innovation, filed September 2019 (project 17042), PCT, Filed 4 September 2020

Abstracts

1. **Hann, E.**, Popescu, I.A., Zhang, Q., Barutçu, A., Neubauer, S., Ferreira, V.M., Piechnik, S.K.. Quality Control-Driven Artificial Intelligence for Reliable Automatic Segmentation of T1 Mapping Images, in: SCMR/ISMRM Co-Provided Workshop, Orlando, Florida, United States. 12 - 15 February 2020. Published: 12 February 2020. **(Thesis Chapter 3)**

2. **Hann, E.**, Ferreira, V.M., Neubauer, S., Piechnik, S.K.. Deep Learning for Fully Automatic Contouring of the Left Ventricle in Cardiac T1 Mapping, in: CMR 2018 – A Joint EuroCMR/SCMR Meeting. Barcelona, Spain. 31 January – 3 February 2018. Published: 31 January 2018.

3. Zhang, Q., Werys, K., Lukaschuk, E., Popescu, I.A., **Hann, E.**, Neubauer, S., Ferreira, V.M., Piechnik, S.K.. Train the Ai like a human observer: deep learning with visualisation and guidance on attention in cardiac T1 mapping, in: Heart. BMJ, pp. A8–A9. BSCMR Annual Meeting 2019, Oxford, England. 27th March 2020. https://doi.org/10.1136/heartjnl-2019-bscmr.9

4. Fung, K., Biasiolli, L., **Hann, E.**, Aung, N., Paiva, J., Lukaschuk, E., Sanghvi, M., Carapella, V., Rayner, J., Werys, K., Puchta, H., Thomas, K., Moon, N., Khanji, M., Neubauer, S., Piechnik, S., Munroe, P.B., Petersen, S.. Effect of coffee consumption on arterial stiffness from UK biobank imaging study, in: Heart. BMJ, p. A8.2-A10. BSCMR Annual Meeting 2019, Oxford, England. 27th March 2020. https://doi.org/10.1136/heartjnl-2019-bcs.9

5. Fung, K., Biasiolli, L., Aung, N., **Hann, E.**, Paiva, J.M., Lukaschuk, E., Sanghvi, M.M., Carapella, V., Rayner, J.J., Werys, K., Thomas, K., Moon, N.O., Neubauer, S., Piechnik, S.K., Petersen, S.E.. Reference values for aortic distensibility derived from UK Biobank cardiovascular magnetic resonance (CMR) imaging cohort, in: Eur. Hear. J. - Cardiovasc. Imaging. EuroCMR 2019, Venice, Italy. 2-4 May 2019. https://doi.org/10.1093/ehjci/jez114

6. Fung, K., Biasiolli, L., **Hann, E.**, Ramirez, J., Lukaschuk, E., Aung, N., Paiva, J., Werys, K., Sanghvi, M., Thomson, R., Rayner, J., Puchta, H., Moon, N., Thomas, K., Lee, A., Piechnik, S., Neubauer, S., Petersen, S., Munroe, P.. First Genome-Wide Association Study of Cardiovascular Magnetic Resonance Derived Aortic Distensibility Reveals 7 Loci, in: Artery Research. Atlantis Press, p. S21. December 2019. https://doi.org/10.2991/artres.k.191224.015

7. Sanghvi, M., Biasiolli, L., Aung, N., Cooper, J.A., Fung, K., Lukaschuk, E., Paiva, J.M., Carapella, V., **Hann, E.**, Rayner, J.J., Werys, K., Puchta, H., Piechnik, S.K., Neubauer, S., Petersen, S.E.. The impact of modifiable cardiovascular risk factors on aortic distensibility: insights from the UK Biobank, in: European Heart Journal - Cardiovascular Imaging — Oxford Academic. 2019. https://doi.org/10.1093/ehjci/jez103

Abbreviations

AA Ascending Aorta

ACC Accuracy

ANOVA Analysis of Variance

AoD Aortic Distensibility

AUC Area Under the Curve

CI Confidence Interval

CMR Cardiac Magnetic Resonance

CNN Convolutional Neural Network

CT Computed Tomography

CVD Cardiovascular Diseases

DSC Dice Similarity Coefficient

FCN Fully Convolutional Neural Network

FN False Negative

FP False Positive

FPR False Positive Rate

GPU Graphical Processing Unit

GT Ground Truth

LGE Late Gadolinium Enhancement

LV Left Ventricle

MAE Mean Absolute Error

MRI Magnetic Resonance Imaging

NN Neural Network

OCMR Oxford Centre for Clinical Magnetic Resonance Research

PDA Proximal Descending Aorta

PRC Precision

REC Recall

ROC Receiver Operating Characterisitic

RF Random Forest

RNN Recurrent Neural Network

RV Right Ventricle

SAX Short Axis

SD Standard Deviation

ShMOLLI Shortened Modified Look-Locker Inversion Recovery

TN True Negative

TP True Positive

TPR True Positive Rate

UKBB UK Biobank

WIP Work In Progress

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cardiovascular Magnetic Resonance Imaging

Cardiovascular diseases (CVDs) are among the leading causes of death worldwide, killing more than 15 million people in 2016 alone [1]. Approximately 10% (7 million) of the UK population have been diagnosed as having some form of CVD [2]. The high risk of mortality signifies the enormous value of investigating these diseases.

Cardiovascular magnetic resonance (CMR) is one of the major non-invasive imaging modalities for comprehensive investigation of the heart in current clinical practice. While CMR scanning can be expensive and may require breath-holds, CMR offers multiple practical advantages compared to other major cardiovascular imaging modalities. Compared to cardiovascular computed tomography, CMR does not use ionising radiation [3]. Moreover, CMR has better reproducibility, and likely better quality control than echocardiography, even without the use of contrast agents [3]. Furthermore, CMR offers non-invasive tissue characterisation for different clinical needs, including T1/T2 weighted imaging, tagging, perfusion, and quantitative T1/T2 mapping etc [4].

CMR is increasingly used in large-scale clinical studies to study various cardiac diseases [3, 5]. For instance, the UK Biobank is a prospective cohort study aiming to recruit 500,000 participants, 100,000 of which will be scanned for the Imaging Component by 2021 [3], with more than 48,000 datasets acquired already. Acquisition of a large number of datasets is necessary to allow reliable studies of particular diseases, as only a relatively small number of cases will match particular conditions [3]. To study cardiovascular diseases, CMR scans are included in the Imaging Component alongside brain MRI, abdomen MRI, and carotid ultrasound and Dual-energy X-ray absorptiometry, with follow-up scans after 5, 10, and 15 years [3]. A UKBB CMR dataset includes but is not limited to long-axis and short-axis cines , aortic distensibility cine, and native T1 mapping [4]. These will contribute to the multi-organ multi-modality imaging resources in the UK Biobank to investigate the mechanism of cardiovascular diseases and beyond [3, 4].

1.2 Automated Medical Image Segmentation

1.2.1 Background

After acquisition of a CMR dataset, segmentation of relevant anatomical structures of interest is usually required to calculate useful clinical parameters, such as ejection fraction, scar burden, and global or regional myocardial T1 values. In the case of CMR imaging, segmentation is also known as contouring, particularly when referring to the delineation of specific anatomic boundaries, such as the endocardial or the epicardial surfaces of the heart.

In current clinical practice, manual segmentation is still the gold standard. This is a tedious, time-consuming and subjective process. In the case of the UK Biobank Imaging component [3], this could potentially require years of manual contouring for a single analyst. While sharing work between multiple analysts can speed up the process, it introduces inter-observer variability, reducing consistency, which may increase the sample size required to detect primary endpoints [6]. Hence, there is a pressing need for processing large-scale CMR datasets consistently and efficiently. To address this need, it is desirable to develop robust, fully-automatic accurate segmentation algorithms for advanced imaging techniques with reliable quality control.

1.2.2 Related Work

There has been extensive research on automating CMR image segmentation since the early 1990s, attempting to address various technical challenges, including anatomical shape variability, suboptimal image quality, presence of papillary muscles, pathologies, partial volume effects, and image artefacts [7, 8]. It has been found that vast majority of the CMR image segmentation methods published from 2000 to 2016 were based on 7 popular techniques: thresholding, region-growing, pixel or voxel classification, active contour, direct estimation, atlas-based segmentation, and statistical shape modelling [8]. Of all the CMR image analysis publications reviewed by [8], only one study reported using deep learning in the methodology, applied to the estimation of bi-ventricular volumes [9].

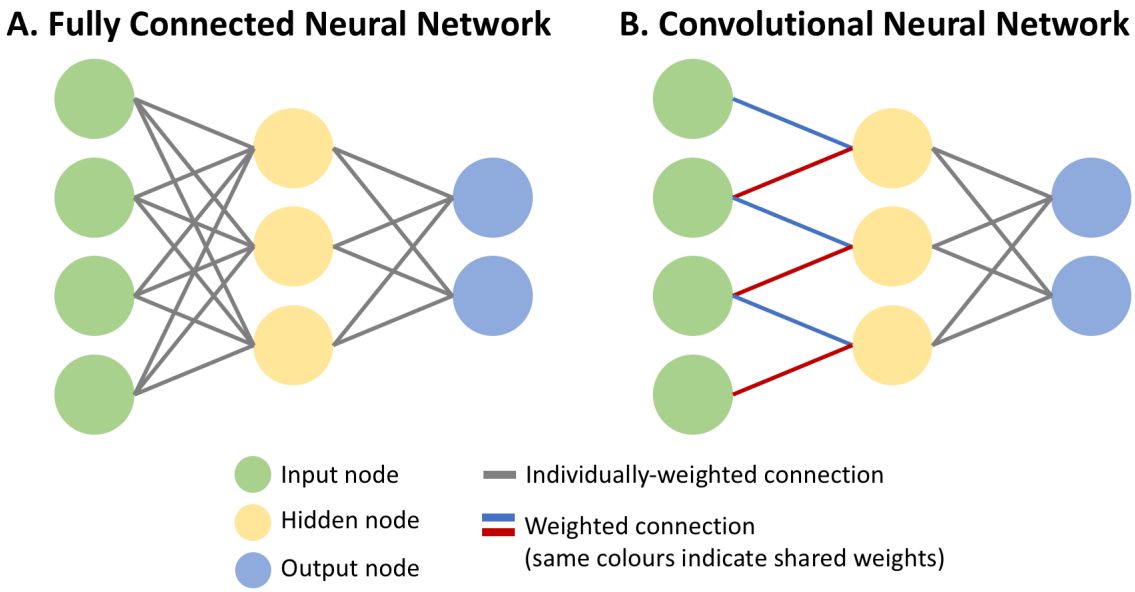


Figure 1.1: Simplified examples of (A) a fully-connected neural network and (B) a convolutional neural network. (A) This is a generic architecture of vanilla fully connected neural network, in which all nodes between layers are connected one-to-one. Each connection carries an individually learnable weight, which can be optimised by training the neural network. (B) This is a generic convolutional neural network, in which connections with shared weights are introduced. This reduces the number of parameters to train, compared with a fully-connected neural network of similar depth.

Over the past few years, deep learning has quickly gained popularity in medical image analysis research, including CMR image segmentation. In 2016, more than 200 deep learning-based research articles were published by the wider medical image analysis community [10]. A similar trend has been observed for CMR image segmentation research with more than 130 papers published on deep learning-based methods since 2016 [11]. The gain in popularity of deep learning in CMR segmentation research can be attributed to the advancement of computing hardware and the increased availability of public datasets [11].

Among deep learning models, convolutional neural networks (CNNs) have been the most successful for image analysis [10]. Figures 1.1A and B, respectively, illustrate a simplified fully-connected neural network and a simplified convolutional neural network. Compared to a standard feed-forward neural network of similar depth, a

CNN has fewer parameters, which are shared among connections at the same level, so that the training is more efficient [12]. This improvement is beneficial for processing large multi-dimensional data, such as images. Note that the last layer of the CNN in Figure 1.1B is fully connected to perform image classification [12].

The medical image analysis research community has drawn inspiration from successes of deep learning in computer vision, such as AlexNet [12], outperforming previous state-of-the-art by a large margin for a large-scale visual recognition challenge [13, 10]. Besides the exceptional performance, another prominent advantage of deep learning CNN models is the ability to learn intricate features for visual recognition tasks solely from data [13, 10, 11]. This relieves the burden of manual feature engineering, which requires handcrafting features for computers to recognise specific structures under various challenging conditions present in the imaging data [7, 8, 13, 10, 11]. Thus, this advantage of deep learning can facilitate translation to various CMR image analysis tasks [11].

For the specific task of CMR image segmentation, one of the early uses of CNNs was implemented in a hybrid approach [14]. A CNN was employed to detect the location of the left ventricular (LV) chamber for short-axis (SAX) cine CMR, prior to a shape initialisation by another neural network, and then segmentation using a deformable model [14]. The output size of the CNN used in [14] was limited to 32 by 32 pixels (versus 256 by 256 pixels for a CMR image), thus more suitable for object localisation than for segmentation.

After that, CMR image segmentation research has progressed towards using fully convolutional neural networks (FCNs) [16], to mitigate the limited resolution problem encountered in [14]. An FCN was implemented to learn end-to-end from input image to ground truth for segmentation of the LV and the right ventricle (RV) in SAX cine CMR, outperforming the previous state-of-the-art [15]. The advantage of the FCN, shown in Figure 1.2, is the introduction of an upsampling layer to provide full resolution segmentation output, without the need to use other segmentation techniques such as deformable models.

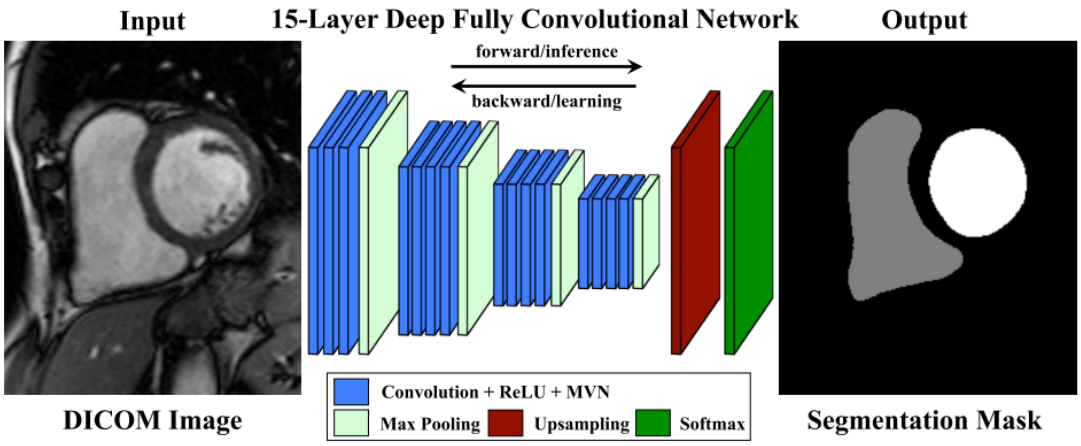


Figure 1.2: A fully convolutional neural network architecture, in which an upsampling layer is introduced to enable segmentation output, with full resolution as the input image. This figure is adapted from [15].

An important development of FCN for biomedical image segmentation is the U-net, originally designed for neuronal structures in electron microscopy stacks [17]. Building upon FCN [16], the U-net (illustrated in Figure 1.3) also employed upsampling layers in addition to convolutional layers for full-resolution segmentation [17]. Skip connections were added to concatenate features from lower layers to higher layers, allowing utilising features of different resolutions to obtain more precise segmentation [17]. U-nets have been used for CMR image segmentation with promising performance [11]. For example, a U-net was deployed to segment the LV myocardium in the T1-weighted images, with over 90% success rate [18].

It has been proposed that overall performance can be improved by employing more than one independently-trained neural network for the same segmentation task. An ensemble of a 2D U-net and a 3D U-net was proposed to segment CMR cine with pathologies [19]. This work demonstrated that averaging the segmentations generated by the ensemble could yield better accuracy than that achieved by either of the individual U-net alone [19]. Another work also achieved good segmentation performance using a similar ensemble of three 2D CNNs and a 3D CNN to perform CMR segmentation, cascaded with another 3D CNN to generate the final segmentation [20]. Thus, using U-nets with ensemble learning is a promising approach for CMR image

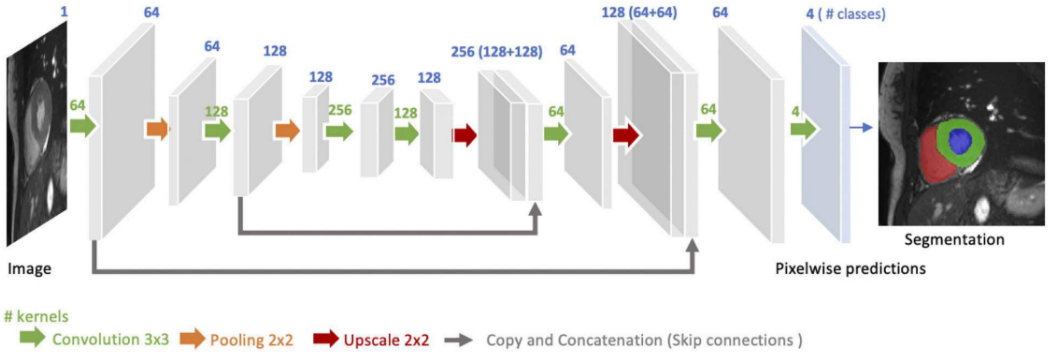


Figure 1.3: A U-net architecture is shown with skip connections (copy and concatenation) from lower layers in the extracting path (left half) to higher layers in the symmetric expanding path (right half), for more precise segmentation. This figure is adapted from [11].

segmentation. Further, [21] used an ensemble of 50 U-nets to estimate uncertainty of the prediction, in addition to the segmentation tasks for brain, heart, and prostate. This has demonstrated a new perspective of segmentation and uncertainty estimation to applying ensemble learning of deep neural networks for CMR image segmentation and inherent quality control.

1.2.3 Summary

In summary, deep learning, particularly convolutional neural networks such as U-net, has recently evolved into a method of choice for image segmentation in medical imaging including CMR imaging. Additionally, the recent development and introduction of ensemble learning show further potential to improve segmentation performance. Ensemble learning of neural networks has shown to estimate uncertainty in addition to image segmentation. This can be further explored for quality control of segmentation results, which is a crucial step especially for clinical applications.

1.3 Quality Control for Medical Image Segmentation

1.3.1 Background

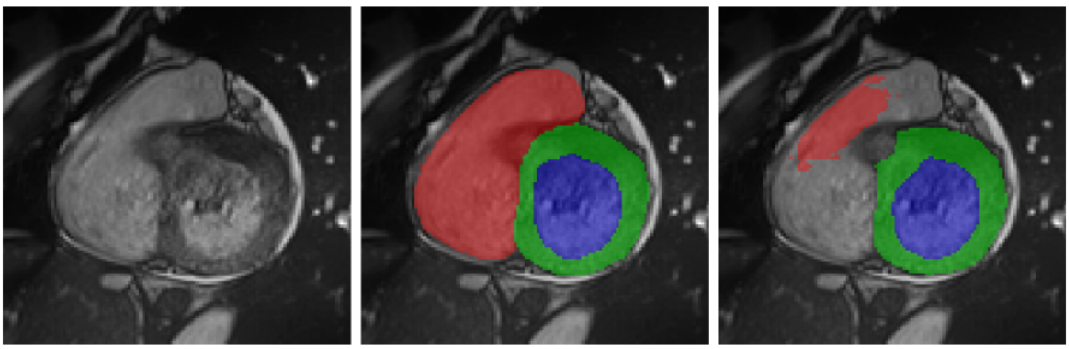


Figure 1.4: Example of a basal short-axis CMR image (left), the ground truth segmentation (middle), and incorrect segmentation by a deep learning-based automation (right). This figure is adapted from [22].

Quality control for medical image segmentation aims to ensure that the segmentation results meet the standard required for research or clinical use. As mentioned in the previous section, medical image segmentation is an essential step to estimate clinical parameters in CMR. Even for manual segmentation, inaccuracy and variability can affect clinical parameter estimation [23, 24]. The same issue is also applicable to automatic segmentation (Figure 1.4). If inaccurate segmentations go undetected, this can lead to incorrect diagnostic or research conclusions, potentially harming the affected patients. Thus, quality control of segmentation is crucial to prevent such undesirable consequences.

Conventionally, segmentation algorithms are evaluated against the ground truth available in a testing dataset to aggregate overall segmentation performance. [22] evaluated the state-of-the-art deep learning-based CMR segmentation methods to measure segmentation performance metrics such as Dice similarity coefficient (DSC). As illustrated in Figure 1.5, DSC is calculated between automated and manual segmentation masks, yielding a score between 0 (no agreement with the manual ground truth) and 1 (perfect agreement with the manual ground truth). However, an aver-

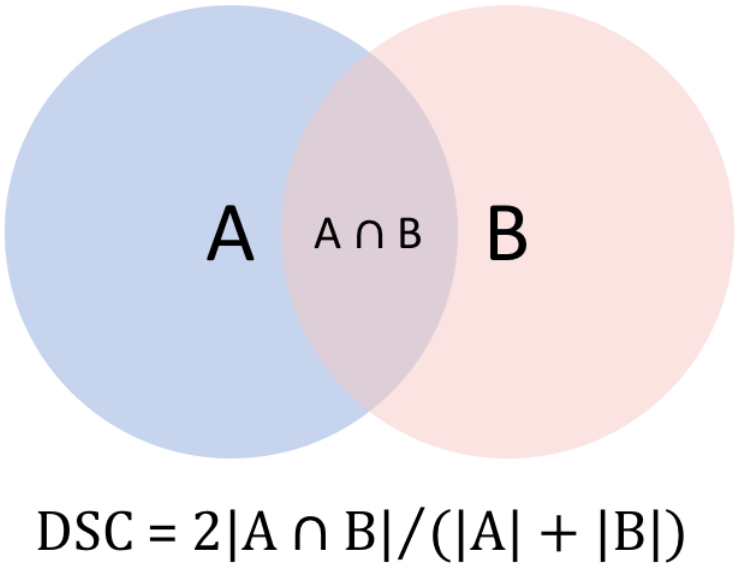


Figure 1.5: Dice similarity coefficient (DSC) is calculated between automated (A) and manual (B) segmentation masks. In essence, DSC measures how much the two segmentation masks overlap, compared to the sum of the areas of the two masks. The score is between 0 (no agreement with the manual ground truth) and 1 (perfect agreement with the manual ground truth).

age DSC does not inform the per-case segmentation correctness to detect individual failures. For example, despite having achieved a high average DSC of over 0.9, the deep learning-based segmentation still failed for some individual cases [22], as shown in Figure 1.4. These failures can potentially affect clinical diagnosis of individual patients, thus it is important to predict the quality of segmentation on a per-case basis, especially for clinical deployment, in which the ground truth segmentation may not be available.

Visual assessment can be performed manually to ensure segmentation quality on a per-case basis. For [25], 2 image analysts performed visual assessment by comparing the automatic CMR cine segmentations to the manual ground truth on a small subset of 250 subjects. Similarly for [26], 13 image analysts, across 2 imaging laboratories, validated automatic localisations of 5100 CMR cine datasets of aortic sections. Scaling up such manual quality control for large databases, such as the complete UK Biobank material, can come with a hefty cost of human labour and

time. Moreover, like manual segmentation, visual assessment is subject to intra-and interobserver variability. Thus, developing automated quality control is important for efficient and reliable deployment of medical image segmentation in large-scale data analysis pipelines and clinical applications.

1.3.2 Related Work

An intuitive approach to automate segmentation quality control is by formulating domain knowledge, such as shape and appearance characteristics, to assess the morphology of image segmentation. Two numeric quality scores were proposed for quality control of myocardial perfusion single-photon emission computed tomography segmentation [27]. The quality scores were formulated based on orientation, area, volume, eccentricity, and intensity of the medical image and the segmentation [27]. Segmentation failures were detected with thresholding on the quality scores. In another study, a similar method was implemented to predict segmentation quality metrics, such as overlap error and DSC, based on 42 shape and appearance features [28]. Similar to [27], segmentation failures were detected with thresholding on the predicted metrics. These methods were limited by whether the shape and appearance features cover a wide enough spectrum to generalise to other datasets.

By modelling appearance features of a test segmentation, an unsupervised approach was proposed to generate an estimated reference segmentation for quality control purpose [29]. DSC was measured between such estimated reference and the test segmentation, in an attempt to correlate with the ground truth DSC, which was measured between the test segmentation and the manual ground truth. The assumption was that the more clearly visible a structure is, the more the estimated reference and the test segmentation resemble each other; otherwise they diverge from each other [29]. However, this method may not be suitable for more challenging tasks, such as segmentation of myocardial scar tissue or apical short-axis slices, in which the structures of interest may not always have clear boundaries.

Alternative to using appearance model [29], estimated reference segmentations can be generated by using an autoencoder, trained with good quality image-segmentation

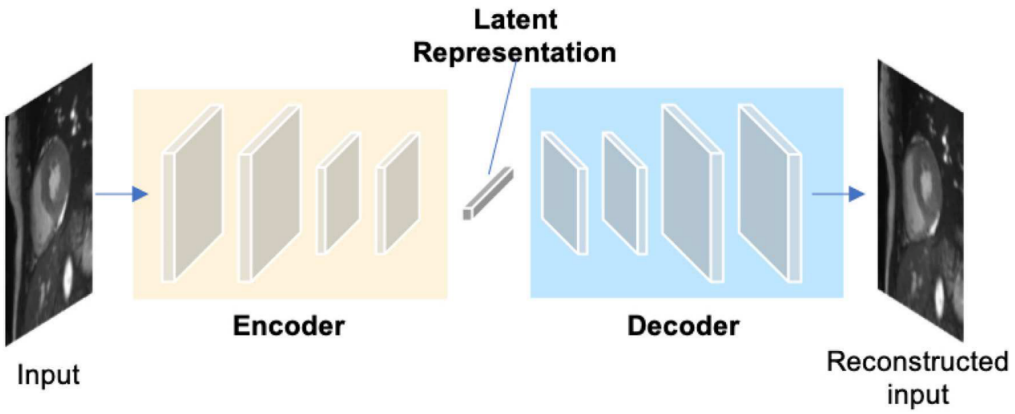


Figure 1.6: A generic autoencoder architecture, comprised of an encoder (left half), a latent representation (middle), and a decoder (right half). Unsupervised autoencoders can learn to encode the input into a low dimensional latent representation, then decode it to a reconstructed input. This is similar to lossy compression of images. This figure is adapted from [11].

pairs [30]. As illustrated in Figure 1.6, an autoencoder can learn to reconstruct its input, by encoding the input to a latent representation and then decoding the latent representation to a reconstructed input. Once trained, an estimated reference segmentation can be generated by the decoder for a given image-segmentation pair. The quality metrics (e.g. DSC) measured between the estimated reference and the test segmentation is output as the predicted quality metrics.

Reverse accuracy classification (RCA) is another approach for segmentation quality control by predicting segmentation quality metrics with estimated segmentation reference [31, 32, 33]. A diverse set of medical images, paired with segmentations, were stored in a reference database. 100 reference medical images were registered to a test image to obtain deformation fields [33]. Based on these deformation fields, 100 corresponding reference segmentations were warped and compared to the test segmentation to calculate 100 DSC values, the maximum of which was output as the predicted DSC for the test segmentation [33]. This approach has been successfully applied to 4800 CMR cine scans available from the UKBB [33]. However, the RCA approach required long computation time (up to 40 minutes per image [33]), making it unsuitable for real-time clinical application.

For real-time application, a convolutional neural network (CNN) was implemented to predict DSC for random forest-based segmentation of CMR cine [34, 35]. The CNN was trained by mapping automatic segmentation masks to ground truth DSCs, generated with corresponding manual segmentations [34, 35]. Once trained, the CNN can make a prediction in a split second (on average 40 ms per image), in the absence of the manual ground truth [34].

Similarly, a CNN was implemented for segmentation of skin lesion images and quality control [36]. The CNN takes an image paired with automatic segmentation and an uncertainty map as input. The uncertainty map served to capture pixel-wise segmentation uncertainty, and could be generated via different techniques such as maximum softmax probability, Monte Carlo dropout, heteroscedastic classifier neural network, and learned confidence estimates [36]. It was demonstrated that incorporation of uncertainty maps could improve segmentation quality prediction [36].

It is also possible to utilise uncertainty information, estimated with the Monte Carlo dropout approach, to predict segmentation quality metrics [37, 38]. Multiple segmentation samples are generated by a U-net incorporated with dropout units, which can randomly null the internal states of the U-net. These segmentation samples would likely differ from one another due to the randomness introduced internally in the U-net by the dropout. Segmentation uncertainty, estimated by measuring the variability among these segmentation samples, was utilised to predict segmentation quality metrics such as DSC [37, 38]. Similarly, [39] used a U-net with Monte Carlo dropout to segment and estimate uncertainty for CMR myocardial arterial spin labelling. [40] compared the Monte Carlo approach with the Bayes by Backprop approach, and obtained similarly good performance from both for cardiac MRI segmentation.

Uncertainty estimation can also be performed with other architectures, such as a probabilistic U-net [41]. Unlike a conventional U-net, the probabilistic U-net aimed to learn and embed distribution of segmentation variants, instead of learning fixed parameters for deterministic segmentation output. Similar to the Monte Carlo approach, the probabilistic U-net could generate a diverse set of segmentation samples to

estimate segmentation uncertainty. Empirically, the probabilistic U-net outperformed other techniques, such as Monte Carlo dropout, for lung abnormalities segmentation and cityscapes segmentation [41].

Similarly, a hierarchical probabilistic model (PHiSeg) could also learn the distribution of segmentation variants with multiple resolution levels [42]. The PHiSeg generated 100 samples for segmentation as well as uncertainty estimation for thoracic tomography (CT) and prostate MRI, with promising results [42].

For CMR native T1 mapping, a CNN was implemented in addition to a PHiSeg [42] to classify segmentation correctness for CMR native T1 mapping [43]. Given an input image, the PHiSeg generated a segmentation and an uncertainty map. Similar to [36], the CNN took the input image, paired with the segmentation and the uncertainty map, and then output a binary classification of segmentation correctness. To train the CNN, 1500 segmentations were annotated manually by an expert for correctness [43]. Like manual segmentation, this manual process to annotate segmentation quality can be time consuming and laborious.

Unlike other Bayesian approaches with segmentation sampling, [44] proposed a calibrated Bayesian network for segmentation with inherent uncertainty estimation, without the need to generate multiple segmentation samples. The authors observed that deep neural networks suffered from miscalibration (overestimation or underestimation) in predicting class probabilities, compared to observed class probability. This problem was mitigated with calibration by introducing an additional utility function [44]. It was demonstrated with application in segmenting brain tumours and thoracic anatomies [44].

Following the success of deep ensembles [45] for image classification tasks, [21] further demonstrated the potential of using deep ensembles for medical image segmentation quality control. [21] used an ensemble of 50 independently-trained U-nets to generate an average segmentation, with calibrated class probabilities, and then to detect out-of-distribution cases. Furthermore, recent research found that ensemble deep neural networks can generate highly diverse predictions, compared to Bayesian neural networks, such as Monte Carlo dropout [46]. High diversity among prediction

samples could be important in capturing uncertainty information [47]. Thus, it is a promising approach for medical image segmentation quality control.

1.3.3 Summary

Proof-of-principle utility of various automatic quality control approaches have been successfully demonstrated for different medical image segmentation tasks. In particular, deep ensembles [45] have shown promise for uncertainty estimation, outperforming other approaches [46]. However, there has been limited exploration of adapting deep ensembles for CMR segmentation quality control. Furthermore, the current state-of-the-art methods have not explored the use of the predicted quality scores as feedback to further improve segmentation accuracy and robustness.

1.4 Overview of Thesis Chapters

The main objective of this thesis is to investigate the use of a novel deep ensemble-based quality control-driven (QCD) framework for segmentation and quality control of CMR imaging data, such as cine and T1 mapping. The thesis is organised as follows:

Chapter 2 introduces the proposed deep ensemble-based QCD segmentation framework. The QCD framework uses multiple U-nets to perform segmentation and quality control of about 5000 CMR aortic cross-section cine datasets from the UKBB. The framework is trained to exploit segmentation agreement among multiple U-nets for DSC prediction for quality control. Once trained, both segmentation and DSC prediction can be executed automatically, in the absence of manual ground truth. The potential of the QCD segmentation framework is shown for future deployment to reliably process large-scale CMR datasets. The content for this chapter follows closely what has been published in [48] and the patent (PCT/GB2020/050249).

In Chapter 3, the QCD framework and two additional variants of the framework were trained and applied to perform segmentation of the LV myocardium in T1 mapping images, available internally from the Oxford Centre for Clinical Magnetic Resonance Research (OCMR). The QCD framework can be successfully applied to different CMR modalities beyond simple aortic cross-sections, e.g, more complex anatomical structures such as the LV myocardium in T1 mapping. Further, the DSC prediction is extended to perform binary segmentation quality classification by using a straightforward thresholding scheme, with excellent accuracy. The content for this chapter follows closely the manuscript content published in [49].

Chapter 4 presents a quantitative evaluation of the QCD framework. The QCD framework trained in Chapter 3 is evaluated with a large-scale external dataset of over 2000 T1 mapping images, available from the UKBB, which have been manually contoured and validated for good to excellent image quality. The QCD framework is shown to be robust when deployed to the external, previously unseen UKBB datasets

for segmentation and quality control. The segmentation quality classification is optimised for the UKBB data by finding an optimal threshold for the DSC prediction.

In Chapter 5, the trained QCD framework performance is evaluated in another unseen set of data from the UKBB – T1-maps with suboptimal image quality that have not been manually contoured. Without manual contours to serve as the ground truth, the QCD automatic segmentation is evaluated qualitatively by visual assessment on a per-case basis. The segmentations generated by the QCD framework are scored by an expert human observer for segmentation quality using a graphical user interface developed in-house. This complements the quantitative evaluation in Chapter 4 by comparing the quality control component of the QCD framework and the opinion of a human expert on a per-case basis. This helps explore limitations of the QCD framework.

Chapter 6 provides a summary and discussion of the future directions to further develop the QCD framework, paving the way towards practical, real-world implementations.

Chapter 2

Quality Control-Driven Framework Towards Reliable Large-Scale Image Segmentation of Aortic Sections in CMR Cine

2.1 Introduction

2.1.1 Aortic Distensibility

Arterial stiffening due to aging can lead to cardiac and vascular diseases affecting the heart as well as the brain and kidneys, which can result in disability or even death [50]. Aortic distensibility (AoD) is a sensitive and specific subclinical biomarker for early detection of arterial stiffening, even in asymptomatic cases, by measuring the bio-elastic function of the aorta using MRI [51]. Furthermore, a study has found that reduced AoD can serve as an independent prognostic predictor for cardiovascular morbidity and mortality among individuals without symptoms of cardiovascular diseases [52]. Hence, AoD can potentially be a useful clinical tool for early identification of asymptomatic individuals who would benefit from appropriate measures to prevent or delay age-related arterial stiffening, reducing risk for potentially lethal consequences [51]. Further studies on AoD have been on-going to investigate cardiovascular risk factors [53] and genome-wide association [54] with the large-scale UK Biobank data [3].

In current clinical practice, measuring AoD requires 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. Large-scale studies can benefit from automated image segmentation, which can provide not only efficient image segmentation, but also improved consistency and objectivity for diagnosis.

However, as discussed, 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 [22], especially in cases affected by poor image quality or pathologies. It is important to detect any critical inaccuracies, which can lead to misdiagnosis or incorrect research conclusion. Current clinical practice of quality control of automatic segmentations

still requires visual inspection, which diminishes the benefits of efficiency of automated segmentation. It would be desirable to integrate automated quality control into fully-automated image analysis pipelines, to efficiently and reliably extract clinical parameters.

2.1.2 Related Work

Fully-automatic aortic image segmentation methods without quality control have been proposed [25, 26]. A recurrent neural network (RNN) in [25] 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 [26] using random forest (RF) localisation of the aorta, with a large-scale (3900 subjects) 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 localisation method can initialise semi-automatic segmentation methods scuh as active contour models [55]. 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 methods to predict Dice similarity coefficient (DSC) have been proposed to address the segmentation quality control in the absence of manual segmentation. [28] proposed an automated quality scoring of segmentation using machine learning with 42 hand-crafted features evaluated against DSC. More recently, a framework based on Reverse Classification Accuracy (RCA) [32, 33] 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. [35] proposed a simple CNN-based method trained to predict the DSC of segmentations generated by RF-based algorithms. Another CNN-based framework [37] 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 optimise segmentation performance.

2.1.3 Contributions

In this work, a novel quality control-driven (QCD) image analysis framework is presented. The QCD framework uses multiple neural networks to integrate segmentation and quality scoring on a per-case basis. The best final segmentation is automatically selected on-the-fly from multiple candidate models, based on accurate DSC predictions, rather than only passive reporting as in [28, 32, 35, 33, 37]. The effectiveness of this QCD approach has been evaluated on a large-scale dataset of aortic cine images from the UKBB imaging study.

2.2 Methods and Material

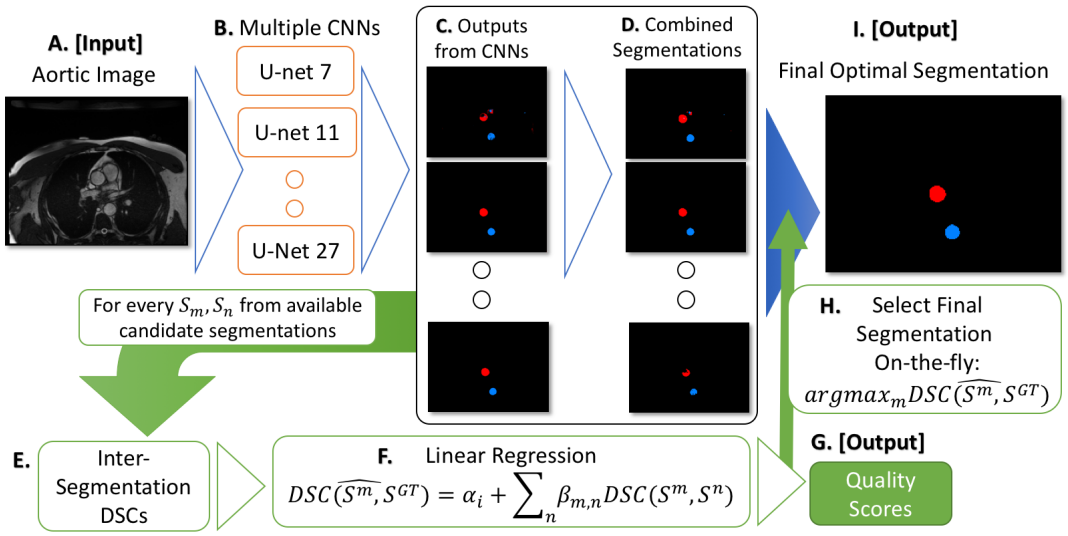


Figure 2.1: The overview of the quality control-driven (QCD) framework, which feeds the same aortic CMR image frame (A) to multiple convolutional neural networks (U-net 7 to U-net 27) (B). Multiple segmentations (C) generated by the U-nets are summed up and thresholded to form additional combined segmentations (D). The inter-segmentation DSCs (E) are 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).

2.2.1 Candidate Segmentation Models

Multiple Convolutional Neural Networks: 6 U-nets [17] with different depths are included as candidate models, for image segmentations of AA and PDA. The 6 U-nets have different numbers of skip connections from 1 to 6 (U-net 7 to U-net 27 in Figure 2.1B). Such differences in the hyperparameters are intended to introduce variation in segmentation performance, which is exploited for segmentation quality control.

Combined Segmentations: A pixel-wise label voting scheme [56] is used to combine multiple U-net segmentations to generate 6 additional segmentations (Figure 2.1D) for improved robustness at a small additional computation cost, such that:

$$C^t(i,j) = \begin{cases} 1, & \text{if } \sum_{U \in \mathbf{U}} U(i,j) \geq t \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

where C^t is a combined segmentation with thresholding parameter t, $U \in \mathbf{U}$ is the segmentation output by a U-net, and (i, j) 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 aortic section.

2.2.2 Quality Scoring and Quality Control-Driven Segmentation

Automatic quality scoring predicts $DSC(S^m, S^{GT})$ for a test segmentation S^m, which is generated by a segmentation model $m \in M$, by comparing with all available candidate segmentations (Figure 2.1C and D) in the absence of the manual ground truth (GT) segmentation S^{GT}. For each segmentation S^m, inter-segmentation DSCs (Figure 2.1E) are calculated with other candidates S^n, generated by another model $n \in M$ and $n \neq m$, of the same input, and then used to predict $DSC(S^m, S^{GT})$ through multiple linear regression $\widehat{DSC(S^m, S^{GT})} = \alpha_m + \sum_{n \in M} \beta_{m,n} DSC(S^m, S^n)$ (Figure 2.1F), where regression parameters α_m and $\beta_{m,n}$ are optimised for each segmentation model m using the training data.

The DSC prediction exploits differences among candidate segmentations. A low DSC is predicted when a test segmentation has low agreement with other candidate segmentations. This could be due to poor image quality of the cine image, causing high uncertainty in segmentation. Thus, it is anticipated that the segmentation may be poor quality. In contrast, a high predicted DSC is expected when there is high agreement among candidates, indicating low segmentation uncertainty. Diversity among candidate models is important to ensure that agreement among candidates can correctly reflect the expected segmentation quality, such that candidates would not produce highly similar erroneous segmentations, which would give a false sense of low uncertainty. It has been found to deep ensemble methods have high predictive diversity among candidate models, thus suitable for uncertainty estimation and quality control [46].

Quality control-driven segmentation uses the DSC prediction to select the best final segmentation (Figure 2.1I). 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 from all candidates on-the-fly: $argmax_{m \in M} \widehat{DSC(S^m, S^{GT})}$ (Figure 2.1H). This is to further improve accuracy and robustness of segmentation by choosing the predicted best on a per-case basis.

2.2.3 Data and Annotations

The dataset comprised of 5028 CMR aortic cine (transverse balanced Steady State Free Precession, with TE = 1.17 ms, TR = 2.8 ms, flip angle = 60 degrees, Grappa factor = 2) image sequences from 4996 subjects (repeated scans were included) acquired under retrospective ECG-gating on a 1.5 Tesla MRI scanner (Model: Siemens Aera, Syngo Platform VD13A) in a single centre (in Cheadle, UK) for the UKBB Imaging Component [4, 26]. In each image sequence, 100 frames across a cardiac cycle were interpolated from the actual temporal resolution of 28 ms, with pixel dimension of 240×196 and resolution of $1.58 \times 1.58 mm^2$ [26].

The manually-validated segmentations of AA and PDA were generated prior to this work using both random forest (RF) localisation [26] and 2D active contour [55]. The RF method selected the most probable AA and PDA locations to initialise the active contour models. The 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 (about 15 out of 100 frames) were manually validated and corrected to reduce the workload on the image analysts. This presented a sparse annotation problem, similar to that reported in [25]. To mitigate this, all generated segmentations were used to train the QCD framework, but only manually-validated segmentations were used for evaluation.

As a standard procedure in machine learning, the manually validated material was randomly split into training data, validation data, and testing data. The training

data were used in the independent learning processes of the candidate U-nets and the DSC regression models. Once trained, all candidate models for segmentation and DSC prediction were evaluated on the unseen validation data. The validation data were used for selecting the best performing segmentation model to proceed to the next step, which involved further evaluation on unseen testing data. This simulated a real-world scenario, in which the validation data used for model selection are different from the target data at deployment. The use of unseen independent testing data for the final evaluation can better reflect the actual performance of the selected model at deployment.

2.2.4 Evaluation

The objectives of the evaluation were 3-fold: (1) to evaluate and compare the segmentation accuracy of all segmentation models, including the QCD segmentation, using DSC; (2) to evaluate and compare 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, which was used to train the U-nets and the DSC regression models. 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.

2.3 Results

2.3.1 Implementation

The framework was implemented in Python, with TensorFlow. Similar to [25], 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 100 cine frames.

2.3.2 Performance of Segmentation Models

Table 2.1: Segmentation performance of the QCD segmentation and each candidate model evaluated with the validation data

	Mean DSC		**Percentage of DSC >0.9**	
Model	**AA**	**PDA**	**AA**	**PDA**
U-net 7	0.918	0.926	77.4	83.7
U-net 11	0.949	0.957	97.5	98.9
U-net 15	0.954	0.961	99.4	99.3
U-net 19	0.951	0.955	98.8	98.7
U-net 23	0.953	0.955	99.4	98.5
U-net 27	0.953	0.956	99.5	99.0
Combined Model 1	0.937	0.942	93.7	92.5
Combined Model 2	0.964	0.964	98.8	99.2
Combined Model 3	0.967	0.966	99.6	99.6
Combined Model 4	0.966	0.966	99.6	99.6
Combined Model 5	0.958	0.962	99.3	99.4
Combined Model 6	0.924	0.934	85.8	90.3
QCD	**0.967**	**0.966**	**99.9**	**99.7**

All segmentation models were evaluated for DSC performance in the validation dataset (Table 2.1). The QCD achieved the highest DSC for AA (0.967) and PDA (0.966) segmentation. Similar segmentation accuracy was also achieved by Combined Model 3, which was selected by the QCD as the best candidate over 60% of the cases. In addition, Combined Models 2-5 obtained higher DSCs than any individual U-nets, showing the benefit of combing multiple neural networks. Moreover, the

results (Table 2.1) showed that the 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. The QCD had the best overall segmentation performance in the validation data.

2.3.3 Quality Scoring of Segmentations

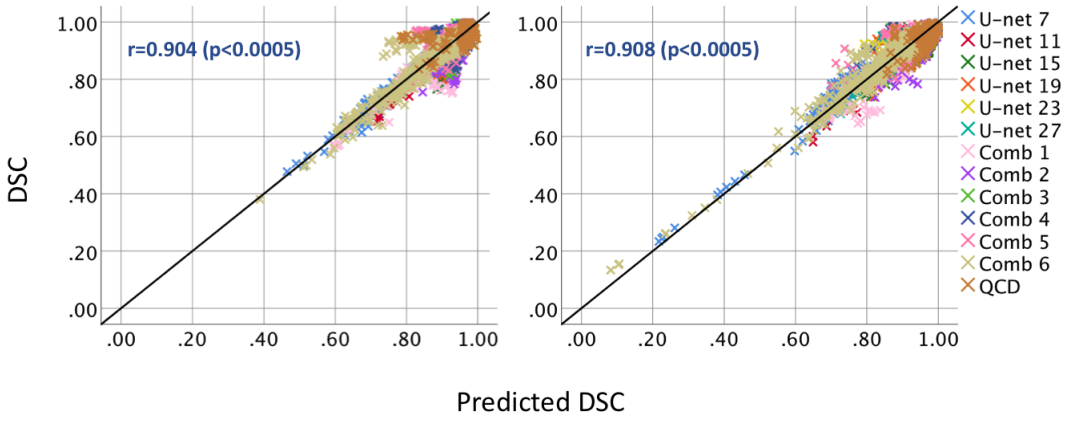


Figure 2.2: Scatter plots of predicated DSC (x-axis) and DSC (y-axis) for AA (left) and PDA (right) in the validation data, with correlation coefficients (r), and p-values for all data points reported. Overall good DSC prediction for all candidate segmentations, with varying quality. Low DSC scores of poor segmentations output by U-net 7 and Combined Model 6 were accurately predicted. Remarks: Comb denotes Combined Model

The segmentation quality scoring was evaluated for all candidate segmentations in the validation dataset. High agreement was achieved 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 correlation coefficients of over 0.9 for both AA and PDA. The scatter plots (Figure 2.2) show that DSC and predicted DSC met along the identity lines, indicating accurate DSC predictions for segmentations of varying quality.

2.3.4 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

Table 2.2: Evaluation results of the QCD framework in the test dataset of 4228 image sequences

Label	Mean DSC	MAE in DSC Prediction	MAE in Lumen Area (mm2)
AA	0.966	0.011	17.6
PDA	0.966	0.015	10.5

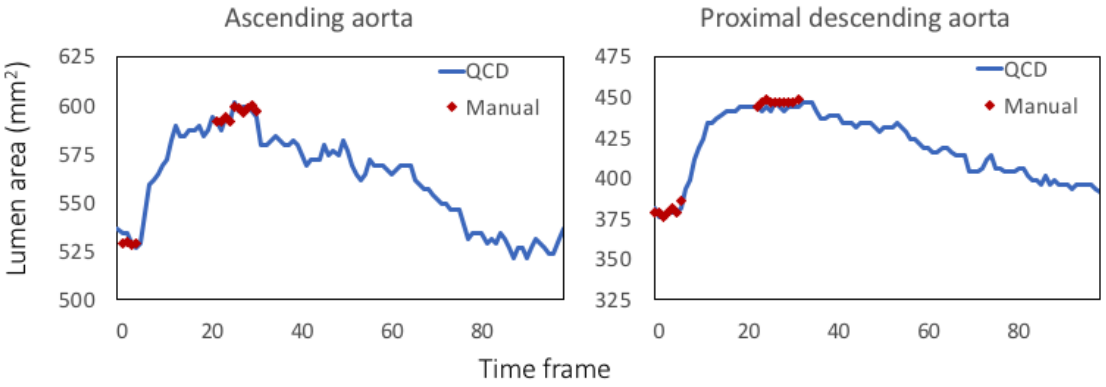


Figure 2.3: Lumen area curves for AA (left) and PDA (right) estimated by QCD (blue), compared with manually validated ground truth (red; only in end-diastolic and end-systolic frames).

performance, with mean DSC of 0.966 for both AA and PDA (Table 2.2), 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 2.2). Two examples of lumen area curves are shown in Figure 2.3. Both curves show consistent lumen area estimation with manual annotations at systole and diastole. In addition, Figure 2.4 shows an example in the testing data to demonstrate how differences in candidate segmentations influence the DSC predictions in the QCD framework.

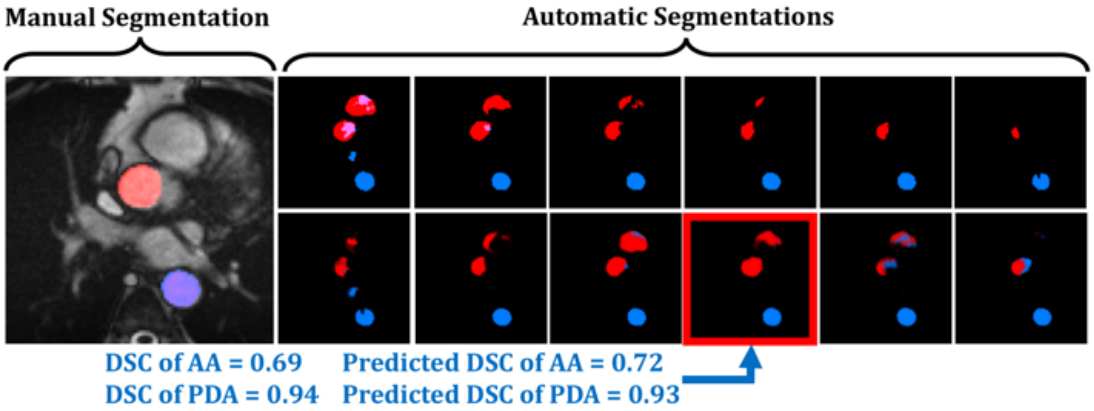


Figure 2.4: Example of a poorly-planned aortic cine image (too far below the main pulmonary artery). Manual segmentation (left large panel), with multiple automatic candidate segmentations of AA (red masks) and PDA (blue masks) are shown. For the final selected segmentation (outlined in red), the predicted 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.

2.4 Discussion

The QCD framework has been successfully developed and evaluated with a large-scale dataset of over 5000 subjects from the UKBB, demonstrating top segmentation and quality control performance. It has been shown that the QCD framework can reliably extract essential clinical parameters (aortic section lumen area in this case) with excellent agreement to the manually estimated ground truth. Together with a high processing speed of 0.67 second per image frame, the QCD framework is suitable to process imaging data for both large-scale research studies and real-time clinical applications.

2.4.1 Comparison with Related Work

Two state-of-the-art automatic segmentation methods have been found for comparison with the QCD framework [25, 26]. For segmentation performance, the two state-of-the-art methods reported mean DSC results of over 0.95, similar to that for the QCD framework [25, 26]. Similar segmentation robustness was also reported in one of the methods with at least 94% of the random forest-based segmentations having achieved DSC above 0.9 [26]. Thus, the QCD framework has achieved excellent segmentation accuracy and robustness, on-par with the state-of-the-art automated methods.

However, it is important to note that even though the QCD framework and these methods were developed and evaluated using the UKBB data, the data were split differently for training and testing. The QCD framework was trained and validated on 800 scans, and tested on 4228 scans; [25] was trained on 400 scans and tested on 100 scans; [26] was trained on 1200 scans and tested on 3900 scans. Therefore, direct comparison should be treated with caution.

2.4.2 Limitations and Future Work

In the current QCD framework, image frames of a cine scan are processed independently by the U-nets used in the QCD framework, without exploiting temporal

information across frames. Given the flexibility of the QCD framework, it is possible to include RNNs as candidate models in addition to the existing U-nets in the future, potentially further improving both accuracy and temporal smoothness of segmentation.

For segmentation quality control, the QCD framework only generates a scalar value of predicted DSC. It is desirable to extend the DSC prediction into segmentation quality classification of good and bad quality. In Chapter 3, such classification of segmentation quality is implemented with a simple thresholding scheme on the predicted DSC.

In this work, the QCD framework has been applied to delineate aortic sections, which are circular in shape. To show adaptability to other imaging applications, it is of high research interest to apply the QCD framework to other imaging techniques and anatomical structures. To achieve this, the QCD framework is applied in the subsequent chapters to segment and quality control the detection of the annular-shaped LV myocardium in CMR native T1 mapping.

2.4.3 Conclusion

In this chapter, a novel quality control-driven segmentation framework comprising of different neural networks has been presented. In the absence of manual annotations, the framework exploits differences among candidate segmentations to predict Dice similarity coefficients, which are then exploited to select the most 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 a 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.

Candidate's Contribution

Conceptualisation, Methodology, Software, Experiments, Analysis, Data curation, Literature reviews, Writing - original draft, Writing - review and editing

Publications

1. **Hann, E.**, Biasiolli, L., Zhang, Q., Popescu, I.A., Werys, K., Lukaschuk, E., Carapella, V., Paiva, J.M., Aung, N., Rayner, J.J., Fung, K., Puchta, H., Sanghvi, M.M., Moon, N.O., Thomas, K.E., Ferreira, V.M., Petersen, S.E., Neubauer, S., Piechnik, S.K.. 2019. Quality Control-Driven Image Segmentation Towards Reliable Automatic Image Analysis in Large-Scale Cardiovascular Magnetic Resonance Aortic Cine Imaging, in: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (Eds.), Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer International Publishing, Cham, pp. 750–758. https://doi.org/10.1007/978-3-030-32245-8_83

2. **Hann, E.**, Piechnik, S.K., Popescu, I.A., Zhang, Q., Werys, K., Ferreira, V.M. "Method and Apparatus for Quality Prediction". Patent application submitted to Oxford University Innovation (project 16045) PCT/GB2020/050249 , Filed 4 February 2020

3. Biasiolli, L., **Hann, E.**, Lukaschuk, E., Carapella, V., Paiva, J.M., Aung, N., Rayner, J.J., Werys, K., Fung, K., Puchta, H., Sanghvi, M.M., Moon, N.O., Thomson, R.J., Thomas, K.E., Robson, M.D., Grau, V., Petersen, S.E., Neubauer, S., Piechnik, S.K.. 2019. 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. https://doi.org/10.1371/journal.pone.0212272

Chapter 3

The QCD Framework for Cardiac T1 Mapping: Adaptability across Imaging Techniques and Anatomical Structures

3.1 Introduction

CMR is one of the major non-invasive imaging modalities for comprehensive investigation of the heart in current clinical practice. In particular, quantitative T1 mapping is an emerging CMR technique for advanced myocardial tissue characterisation on a pixel-by-pixel level [57, 58], and can detect disease beyond conventional CMR methods, such as late gadolinium enhancement (LGE) imaging. T1 mapping is designated as one of the six most innovative imaging methods for evaluating patients with heart failure by the European Society of Cardiology Heart Failure Association [59]. CMR T1 mapping is increasingly used in large-scale clinical studies [3, 5] to study various cardiac diseases, including the UK Biobank Imaging Component [3].

In current practice, extraction of useful clinical parameters, such as the average myocardial T1 value from CMR T1 maps, requires manual segmentation of the left ventricular (LV) myocardium, which is a tedious, time-consuming and subjective process. Thus, developing fully automatic image analysis algorithms, with high efficiency and reliability, can alleviate the burden for manually processing CMR T1 maps.

3.1.1 Related Work

For cardiac T1 mapping, there is limited published literature on automatic segmentation. A non-machine learning approach was recently proposed for automatic LV segmentation and regional analysis of myocardial native T1 values [60]. However, it was developed and validated only on a small cohort of healthy controls (10 subjects), which does not capture the wide range of image variability in larger databases of normal and pathological cases commonly encountered in real-world clinical practice. Another approach relied on a fully-convolutional neural network method to segment T1 weighted images to reconstruct myocardial T1 maps [18]. However, these two methods had no automatic segmentation quality control. Based on probabilistic neural networks, an approach was proposed to contour myocardial T1 maps, with segmentation quality control via an additional convolutional neural network, trained and tested using 1500 manual annotations of segmentation correctness [43]. However,

curation for such manual annotations can be time-consuming and labourious, thus not desirable for large-scale development and evaluation.

3.1.2 Contributions

To promote reliable fully-automated image analysis, I presented a quality control-driven (QCD) segmentation framework in Chapter 2, for delineating aortic sections in CMR cine. In this chapter, the QCD framework is applied to left ventricular segmentation of T1-mapping images, to demonstrate that the framework can be easily adapted for another imaging technique and another anatomical structure. The novel contributions of this chapter include the adaptability of the QCD framework to:

1. Segment a substantially different and more complex anatomical structure (the doughnut-shaped left ventricular myocardium in short-axis), compared to simple circular cross-sections of the aorta in [48]. This is then generalisable to other common forms of cardiovascular imaging, such as echocardiography and cardiac computed tomography, where segmentation of the left ventricular myocardium is also commonly performed.

2. Tailor to a completely different CMR imaging protocol (quantitative mapping) from traditional cine imaging in [48], in terms of MR methodology, imaging parameters, types of artefacts, and clinical purposes.

3. Further validate improvement of segmentation accuracy on-the-fly, selecting the most optimal LV segmentation from multiple candidates based on predicted accuracy. This concept is novel to automatic segmentation and quality control in diagnostic imaging, requiring deeper validation for various applications.

4. Include a visualisation tool for segmentation agreement (novel in this work), to provide visual insights into the traditional "black-box" nature of deep-learning-based image processing, with traceability into the segmentation quality control process.

5. Additionally, we highlight a potential flaw of the Pearson correlation, commonly used as a metric for segmentation accuracy prediction. The Pearson correlation between predicted and actual observed DSCs is dependent on the performance of the segmentation method. It can be paradoxically worse for a better-performing method, and thus is not always suitable for evaluating quality prediction.

3.2 Material and Methods

In this section, the origin of the data used in the development and testing of the novel QCD framework is first described. Then, the methodology is introduced for both the segmentation component and the quality control component, with segmentation quality visualisation. Details of implementation and evaluation of the QCD framework are presented.

3.2.1 Material

The development and testing data comprised of 2383 CMR native (pre-contrast) T1 maps using the shortened modified Look-Locker inversion recovery (ShMOLLI) T1-mapping method [61], acquired on both 1.5 and 3 Tesla MRI scanners (Siemens Avento and Trio, Germany), with a 16 or 32 channel phased-array chest coil, TR/TE = 201.32/1.07 ms, flip angle = 35 degrees, 107 phase encoding steps, interpolated voxel size = $0.9 \times 0.9 \times 0.8\ mm^3$, for our prior research studies [62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77]. The original image dimension ranged from 212 to 384 for height and from 196 to 384 for width, prior to the preprocessing step of zero-padding to 384×384 pixels. All T1 maps were short-axis views of the LV myocardium, varying from basal to very apical slices, with endo-and epicardial contours manually segmented. The manual contours served as the ground truth (GT) segmentation for evaluating automatic segmentation and for deriving the reference DSC to train and test the automatic segmentation quality predictors. The data were randomly split into 80% training data, 9% validation data, and 11% testing data.

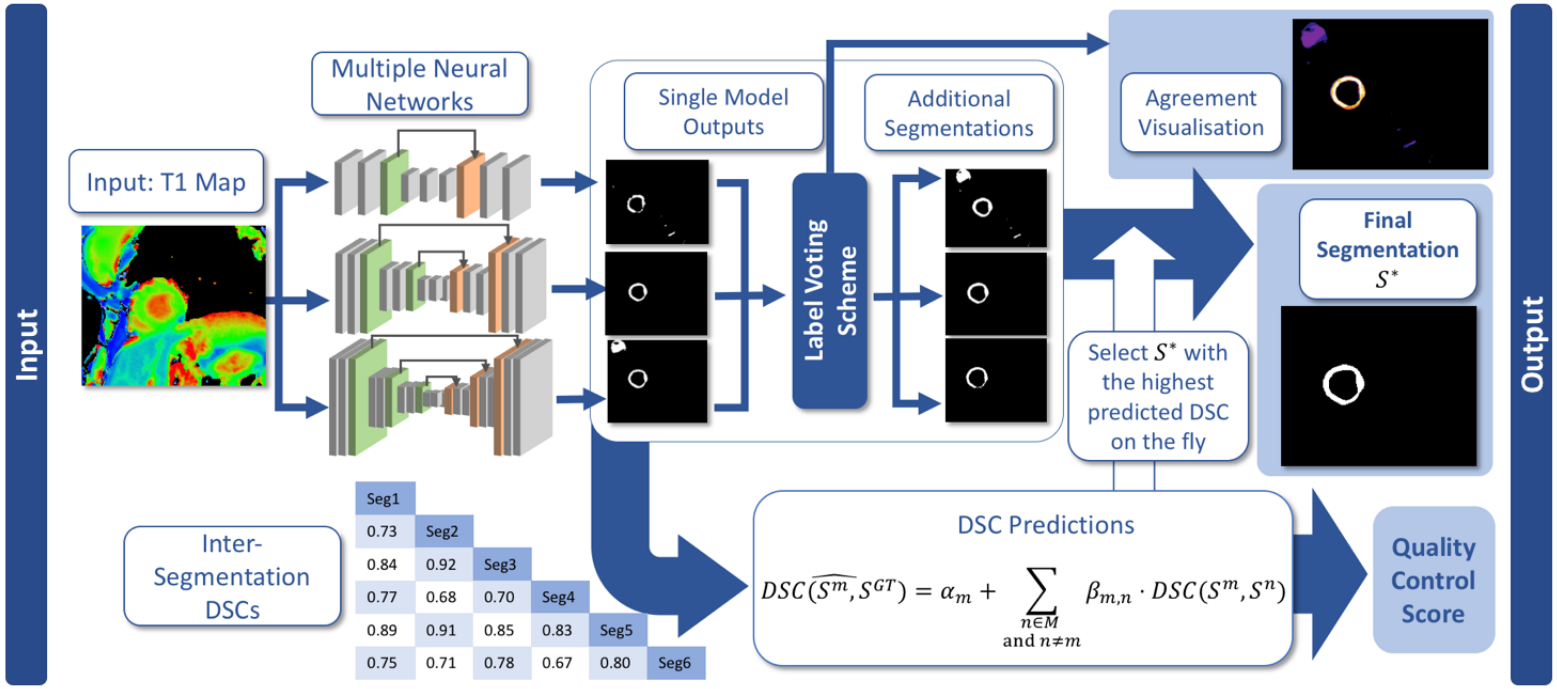


Figure 3.1: Overview of the multiple neural network framework for integrated segmentation and quality control. It follows the same design presented in Figure 2.1 in Chapter 2, with a new feature for visualising segmentation agreement. For simplicity, this illustration shows an example of 3 single independent neural networks. (A) A T1 map is analysed by (B) independent segmentation models to output (C) single-model segmentations. Then, the single model segmentations are passed to a label voting scheme to generate (D) combined model segmentations. (E) In addition, the agreement of the segmentation can be visualised. This is a new feature introduced in this chapter. (F) A DSC matrix is generated from both single model and combined model segmentations for (G) DSC predictions with regression models. (H) The final segmentation is chosen based on the DSC prediction, and the corresponding predicted DSC is output as (I) the final quality control score.

3.2.2 Multiple Neural Network Models

The QCD framework uses multiple segmentation models, where each model $m \in M$ generates a segmentation S^m of an input T1 map (Figure 3.1A). S^m is a binary pixel-classification mask where the LV myocardium is labeled as 1, and other pixels as 0.

There are two types of segmentation models in the framework: single models (Figure 3.1C) and combined models (Figure 3.1D). For an input T1 map, each single model, such as a single convolutional neural network, can independently generate a segmentation (Figure 3.1B). In this work, a range of fully convolutional neural networks of different depths, such as U-net 7, U-net 11, and so on, are used to make a diverse set of candidate segmentations. This is analogous to the spread of expertise in a multidisciplinary clinical team. Furthermore, these single model segmentations can also be combined via a label voting scheme [56] to generate additional segmentation candidates, which we term combined segmentations. All available single model segmentations, denoted as $\mathbf{U}$, of an input T1 map are summed up in a pixel-wise fashion, then thresholded by $t \in \{1, 2, \ldots, \mid \mathbf{U} \mid\}$ such that

$$C^t(i,j) = \begin{cases} 1, & \text{if } \sum_{U \in \mathbf{U}} U(i,j) \geq t \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

where (i, j) is a pixel coordinate in the T1 map, and C^t denotes a combined segmentation generated with a threshold parameter t. This generates $\mid \mathbf{U} \mid$ (the number of neural networks used) additional segmentation variants for each input image.

3.2.3 Visualisation of Segmentation Agreement

The agreement of the single neural network model segmentations is visualised by colour-coding the pixel-wise summation map $\sum_{U \in \mathbf{U}} U(i,j)$ in Eq. 3.1. It highlights the degree and location of segmentation differences among single neural network models (Figure 3.1E), and unmasks the "black-box" nature of the deep learning-based segmentation, facilitating transparency of the quality control process in the framework.

In addition, as combined segmentations are generated similarly by overlaying the single model segmentations pixel-by-pixel, the visualisation also shows the agreement of the combined segmentations.

3.2.4 Automatic Quality Control of Segmentation

In addition to fully-automatic segmentation, the framework is capable of generating an inherent quality score of any segmentation S^m produced by a model $m \in M$, in the absence of the manual ground truth (GT) segmentation S^{GT}. M denotes all the available single and combined models in the framework. For any segmentation S^m, the framework predicts Dice similarity coefficient $DSC(S^m, S^{GT})$ as the segmentation quality score (Figure 3.1G).

The quality scoring exploits the differences in segmentation among all available candidate segmentation outputs to generate the quality score. The quality scoring relies on a negative relationship between the segmentation differences and the segmentation quality.

In order to establish this relationship, the differences in segmentation among the multiple segmentation models, implemented in M, are quantified and compared. $DSC(S^m, S^n)$ is computed for every pair of distinct models $(m, n) \in M \times M$ and $m \neq n$. Hence, inter-segmentation Dice similarity coefficients are calculated (Figure 3.1F) from all the available segmentations in the framework for an input T1 map.

Subsequently, for each segmentation model $m \in M$, a quality scoring model is needed to predict $DSC(S^m, S^{GT})$ of any image. The Dice similarity coefficient prediction $\widehat{DSC(S^m, S^{GT})}$ is based on multiple linear regression, such that:

$$\widehat{DSC(S^m, S^{GT})} = \alpha_m + \sum_{n \in M} \beta_{m,n} \cdot DSC(S^m, S^n), \tag{3.2}$$

where α_m and $\beta_{m,n}$ are the linear regression parameters, trained individually for each segmentation model $m \in M$ using the training data, where the ground truth manual segmentation S^{GT} is available to compute $DSC(S^m, S^{GT})$.

3.2.5 Quality Control-Driven Segmentation

The availability of quality prediction for each candidate segmentation in the framework enables on-the-fly selection of the final segmentation from all the available segmentations. For a T1 map, the segmentation S^m generated by a model $m \in M$ is automatically assigned a quality score, in the form of a predicted Dice similarity coefficient $\widehat{DSC(S^m, S^{GT})}$. Assuming that the predicted Dice coefficient (Figure 3.1G) is accurate, the segmentation S^m with a higher $\widehat{DSC(S^m, S^{GT})}$ is expected to achieve a higher $DSC(S^m, S^{GT})$. Hence, the QCD framework selects the segmentation with the highest quality score $max_{m \in M} \widehat{DSC(S^m, S^{GT})}$ to be the final, most optimal segmentation S^*, for each T1 map (Figure 3.1H). It is expected that this novel QCD approach can improve the overall segmentation accuracy.

Two additional variants of the QCD segmentation have been considered for comparison. The default QCD framework includes both single models and combined models as candidates. The final segmentation is selected based on the highest predicted DSC. The first variant (QCD-Lite) is similar to the default QCD framework. The only difference is that the combined models are excluded from the candidates for the QCD-Lite. This creates a "lighter" version of the default QCD framework. The same independently-trained single models from the default QCD are used as candidates in the QCD-Lite. The DSC predictors are retrained to accommodate fewer candidate models. This is a preliminary attempt to assess how the choice of candidate models impacts on the segmentation performance.

Extending upon the default QCD framework, the second variant (the weighted average QCD) assigns the corresponding predicted DSC as a weight to each candidate segmentation. It then outputs a weighted average segmentation as the final output, instead of selecting only one optimal segmentation. The DSC prediction for the final segmentation is also a weighted average. This is to explore the possibility of further improving the QCD framework.

Table 3.1: Image quality categories for T1 maps described by expert human operators

Category	Description	Proportion
Excellent	Well-defined borders of myocardium with good contrast. Typically, mid-ventricular slice. Easy to contour with high consistency.	5.2%
Good	Overall well-defined borders of the myocardium with reasonably good contrast. Requires some caution when contouring. Moderately easy to contour, but prone to higher variability than easy cases.	23.5%
Acceptable	Ambiguous borders of the myocardium with poor contrast. Requires caution when contouring. Prone to high variability.	65.3%
Poor	Ambiguous borders of the myocardium with poor contrast. Observable pathologies or artefacts.	6.0%

3.2.6 Implementation

For the specific implementation of the QCD framework, 6 independent U-nets [17] were implemented to perform automated LV myocardium segmentation. Each of them varied in hyper-parameters, such as the number of convolutional layers, pooling layers, and the number of skip connections. The smallest neural network implemented had only 7 convolutional layers, and 1 skip connection, while the deepest neural network had 27 layers and 6 skip connections. Each of the neural networks is referred to by the number of convolutional layers as follows: U-net 7, U-net 11, and so forth, up to U-net 27. The wide range in capacity of the networks is intentional to introduce more diverse variation in segmentation. The neural networks were independently trained, using the Adam optimiser [78] to minimise the cross-entropy loss in the training data of CMR T1 maps. The framework was trained and validated on a single desktop computer using a single NVIDIA Titan X GPU, with 12GB onboard memory and 3072 cores. Each convolutional neural network of the ensemble was independently trained for 60 epochs.

3.2.7 Evaluation Methods

For each model $m \in M$, the segmentation performance was evaluated by averaging $DSC(S^m, S^{GT})$ between the automated segmentation S^m and the manual segmentation S^{GT} of T1 maps in the validation data. The accuracy of the DSC prediction was also evaluated using the validation data by mean absolute error (MAE) and Pearson correlation coefficient (r) of the prediction $\widehat{DSC(S^m, S^{GT})}$ and the prediction target $DSC(S^m, S^{GT})$ for each model $m \in M$.

The DSC prediction was further evaluated for binary classification of good (observed DSC ≥ 0.7) and poor (observed DSC < 0.7) segmentation. The threshold of 0.7 was chosen based on [33]. The binary classification was evaluated according to the accuracy $(TP+TN)/(TP+FP+TN+FN)$, the true positive rate $TP/(TP+FN)$, and the false positive rate $FP/(FP+TN)$, where TP, FP, TN, and FN respectively denote the number of true positive cases (observed DSC ≥ 0.7 and predicted DSC ≥ 0.7), false positive cases (observed DSC < 0.7 and predicted DSC ≥ 0.7), true negative cases (observed DSC < 0.7 and predicted DSC < 0.7), and false negative cases (observed DSC ≥ 0.7 and predicted DSC < 0.7). The binary classification can further demonstrate the practical usage of the DSC prediction in the QCD framework.

The estimated myocardial T1 value, calculated by averaging the T1 values of all pixels in the myocardium, was identified by the automated method, for each T1 map in the testing data. Similarly, we established the ground truth T1 value using the manual segmentation. The T1 estimation was evaluated using mean error, mean absolute error (MAE), and Pearson correlation (r) between the estimated values and the ground truth. In addition, the relative errors of T1 were categorised by manual image quality assessments by a consultant cardiologist (AB), who classified the T1 maps into 4 levels of quality: 'excellent', 'good', 'acceptable', and 'poor' (Table 3.1).

Table 3.2: Segmentation and DSC prediction performance for the QCD and candidate models.

Segmentation Model	Mean DSC (SD)	MAE	r
U-net 7	0.669 (0.172)	0.031	0.95
U-net 11	0.809 (0.074)	0.037	0.73
U-net 15	0.830 (0.068)	0.038	0.66
U-net 19	0.826 (0.055)	0.038	0.43
U-net 23	0.831 (0.056)	0.038	0.57
U-net 27	0.831 (0.058)	0.040	0.42
Combined Model 1	0.790 (0.077)	0.046	0.60
Combined Model 2	0.837 (0.060)	0.040	0.47
Combined Model 3	0.837 (0.057)	0.038	0.52
Combined Model 4	0.829 (0.058)	0.035	0.73
Combined Model 5	0.809 (0.073)	0.033	0.88
Combined Model 6	0.688 (0.178)	0.034	0.96
QCD	0.851 (0.054)	0.034	0.53
QCD-Lite	0.850 (0.056)	0.034	0.58
Weighted Average QCD	0.823 (0.059)	0.032	0.71

Remarks: SD = standard deviation, MAE = mean absolute error. All Pearson correlations (r) had $p < 0.0005$.

3.3 Results

The neural networks were trained on 1906 CMR T1 maps and were subsequently validated and tested on previously unseen data of 477 T1 maps. With a single GPU, the framework took 15 minutes and 21 seconds (including data I/O time) to segment the entire dataset of 2383 T1-maps and produce the quality control scores. On average, one image took 0.39 second to process.

3.3.1 Accuracy of Segmentation

Among the 12 individual segmentation models investigated for the QCD framework, Combined Model 3 had the highest mean observed DSC of 0.837 (Table 3.2), followed closely by Combined Model 2 (DSC=0.837), both outperforming the deepest single neural network U-net 27 (DSC=0.831)

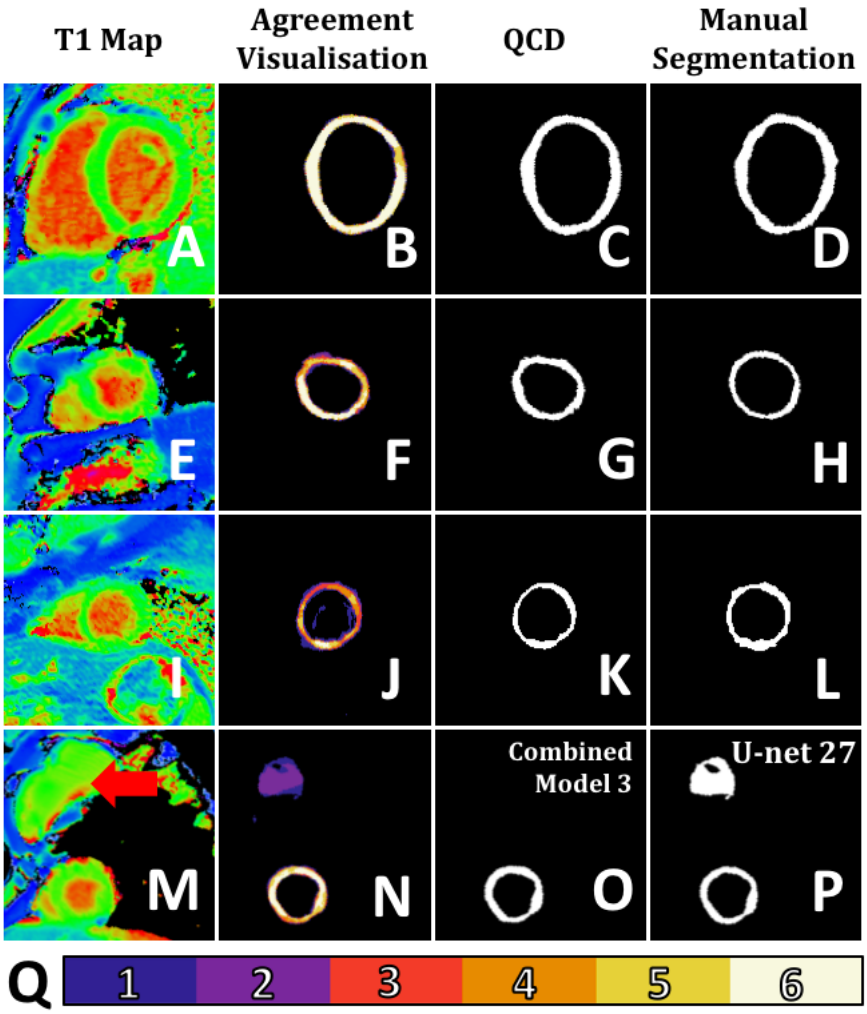


Figure 3.2: Examples of T1 maps, agreement visualisations, and segmentations. (A-D) The top row is an example in which there was high agreement among segmentation models, as shown in (B) the agreement visualisation. Hence, the predicted DSC of the QCD output (C) was high (0.893), which was consistent with the DSC (0.900). (E-H) The second row is an example in which there was some disagreement among the segmentation models, as shown in (F) the agreement visualisation. Hence, the predicted DSC of the QCD output (G) was low (0.655), which was consistent with the DSC (0.643). (I-L) The third row is an example in which the agreement visualisation (J) showed high disagreement among the segmentation models, possibly due to the heavy wraparound artefact. The predicted DSC was low (0.540) due to the disagreement despite that the DSC was much higher (0.791). In clinical practice, this T1 map (I) should be treated with caution. Thus, a lower predicted DSC can serve as a useful alert. (M-P) The last row shows an example in which (P) the deepest single neural network (U-net 27) falsely classified the breast implant (red arrow in M) as part of the myocardium. On the other hand, (O) Combined Model 3 produced more robust segmentation. (Q) is a colour bar which indicates the degree of agreement in the visualisations, with 1 being the lowest agreement to 6 being the highest agreement.

Pictorial examples of the T1 maps and their corresponding segmentations can be seen in Figure 3.2. Specifically, Figure 3.2M-P shows an example that Combined Model 3 generated more robust segmentation than U-net 27. In this case, U-net 27 misclassified the breast implant (indicated by a red arrow in Figure 3.2M) as the myocardium. This case demonstrated the advantage of the on-the-fly selection of the final segmentation combined with the label voting approach, instead of using a fixed segmentation model or a fixed weighted-average segmentation.

Two sets of example candidate segmentations are shown for an easy, good-quality T1 map (Figure 3.3) and a difficult T1 map affected by an extracardiac structure (breast implant) (Figure 3.4). These examples illustrate the agreement among the candidate segmentations under 2 different scenarios. With a good quality T1 map, Figure 3.3 shows high agreement among the candidates with high DSCs of ≥ 0.83. In contrast, Figure 3.4 shows high disagreement, as some of the candidate segmentations failed differently, including falsely identifying the breast implant as the myocardium, and failures to segment the whole myocardium. These demonstrate that segmentation differences from a diverse set of candidates can be exploited to estimate segmentation quality.

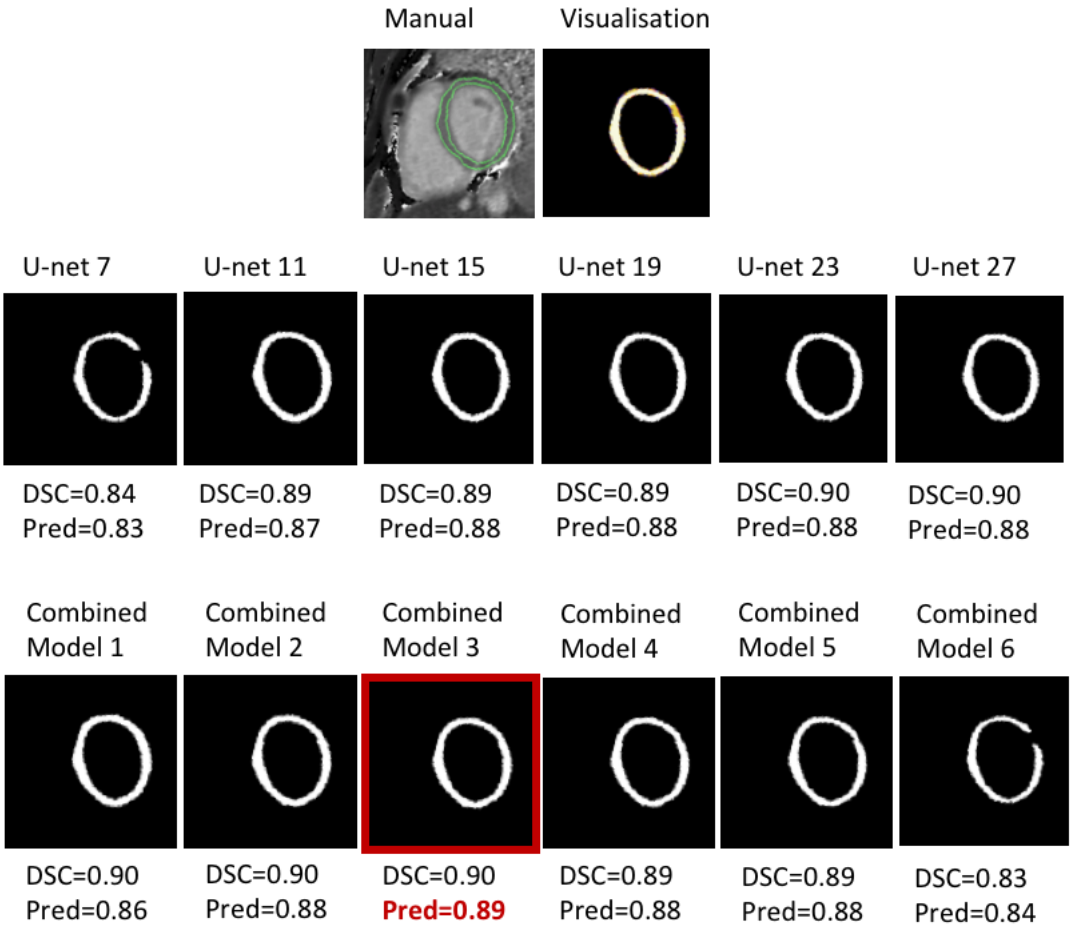


Figure 3.3: Extended example of Figure 3.2A-D showing high agreement among candidate segmentations for a good quality T1 map. The top left image shows the manual segmentation and the top right image shows the visualisation of the candidate segmentations. The rest shows the candidate segmentations from U-net 7 to U-net 27, and the combined segmentations (Model 1 – Model 6). All the candidate segmentations consistently obtained high DSCs of ≥ 0.83, as the good quality T1 map was easy to segment. Combined Model 3 (in the red box) achieved the highest predicted DSC (0.89), thus its segmentation was selected as the final output by the QCD framework.

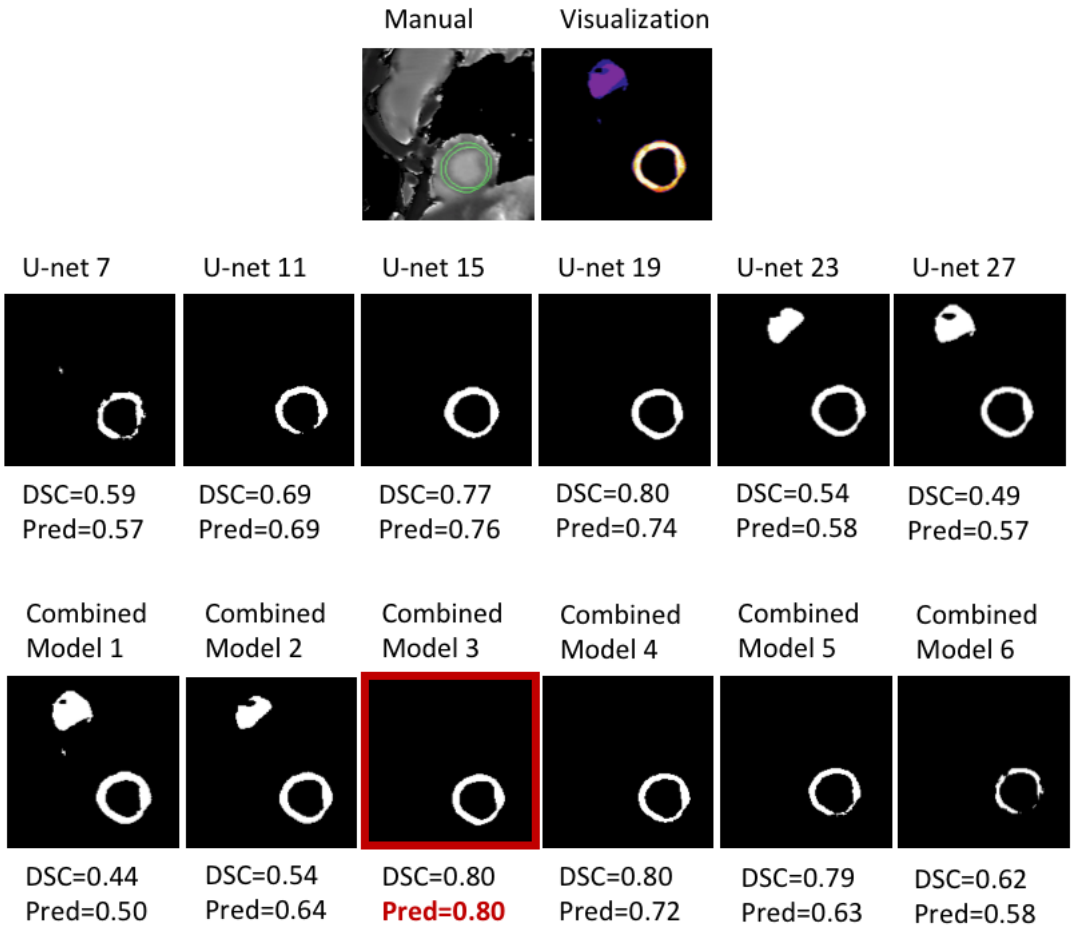


Figure 3.4: Extended example of Figure 3.2M-P showing poor agreement in the candidate segmentation failures of a T1-map affected by an extracardiac structure (breast implant). The top left image shows the manual segmentation and the top right image shows the visualisation of the candidate segmentations. The rest shows the candidate segmentations from U-net 7 to U-net 27, and the combined segmentations (Model 1 – Model 6). When the candidate U-nets failed, they appeared to fail differently, as demonstrated by U-net 7, 11, 23, and 27, obtaining a DSC of 0.59, 0.68, 0.54, and 0.49, respectively. U-nets 7 and 11 failed to form an annulus-like myocardial mask, whereas U-nets 23 and 27 falsely identified the breast implant as part of the myocardium. Despite the difficulty, U-nets 15 and 19, and Combined Models 3 and 4 successfully segmented the myocardium, with DSCs ≥ 0.77. This illustrates the importance of including a diverse set of candidate models. Combined Model 3 (in the red box) achieved the highest predicted DSC (0.80), thus its segmentation was selected as the final output by the QCD framework.

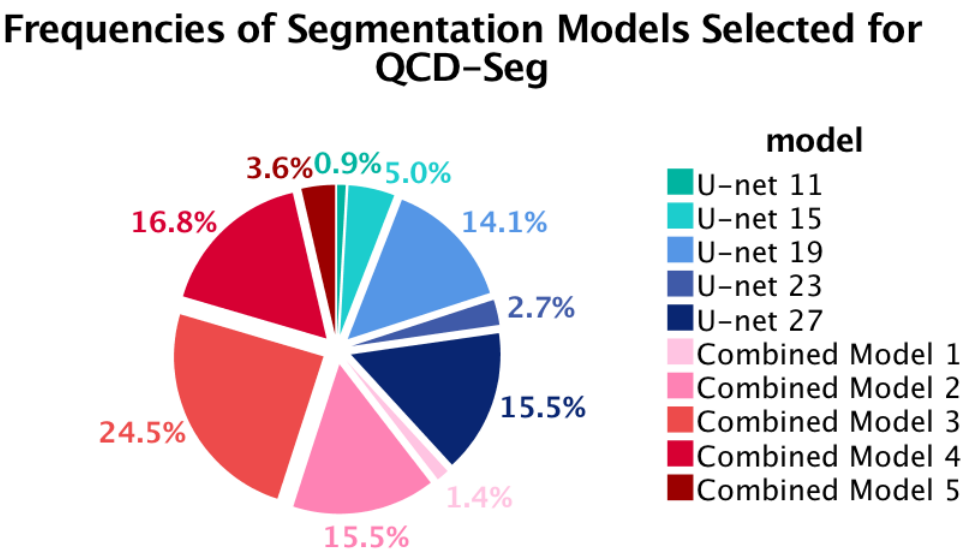


Figure 3.5: Pie chart of frequencies of the segmentation models selected for the final segmentation in the QCD framework. It shows that outputs generated by Combined Models 2, 3, 4 were most frequently selected as the optimal segmentations, accounting for more than half of the cases in the validation data. No segmentation generated by U-net 7 or Combined Model 6 was selected by the QCD framework.

The QCD framework and the QCD-Lite variant further outperformed any individual segmentation models and demonstrated the best performance in the LV myocardium segmentation on the validation data, with a DSC value of 0.851 and 0.850, respectively (Table 3.2). The QCD framework and the QCD-Lite also outperformed the weighted average QCD variant, which obtained a DSC of 0.823. This demonstrated the effectiveness of the optimal segmentation model selection with the highest predicted quality obtained on-the-fly. This is similar to our clinical experience that averaging may fall short when multiple human analysts have different training and experiences. The segmentations produced may not form linear relationships. Combined Models 2, 3 and 4 contributed the most to the QCD segmentation, accounting for more than half of the final segmentation outputs (Figure 3.5).

3.3.2 Visualisation of Segmentation Agreement

The agreement visualisation of segmentation shows a spatial map of agreement among the multiple single neural networks. Additional examples of the agreement visualisation can be seen in Figure 3.2. Figure 3.2Q is the colour bar of scale from 1 to 6, indicating the number of single neural networks which identify a particular pixel as the myocardium, hence showing the extent of agreement among the neural networks. Figure 3.2B shows an agreement visualisation with generally high degree of segmentation agreement across the myocardium segmentation. Thus, the automated segmentation (Figure 3.2C) was also expected to highly agree with the manual segmentation (Figure 3.2D). Figure 3.2F shows that the neural networks disagreed with each other mostly at the apical anterior wall. This is the same region where the automated segmentation (Figure 3.2G) differed from the manual segmentation (Figure 3.2H). Figure 3.2J shows generally high disagreement among the neural networks across the myocardium, possibly due to the heavy wraparound artefact in the T1 map (Figure 3.2I). Thus, a low predicted DSC was expected. Figure 3.2N shows a high disagreement at the breast implant (purple-coloured pixels). These examples show that the agreement visualisation can highlight the regions where disagreements happen and provide insights into the quality control of the segmentation process.

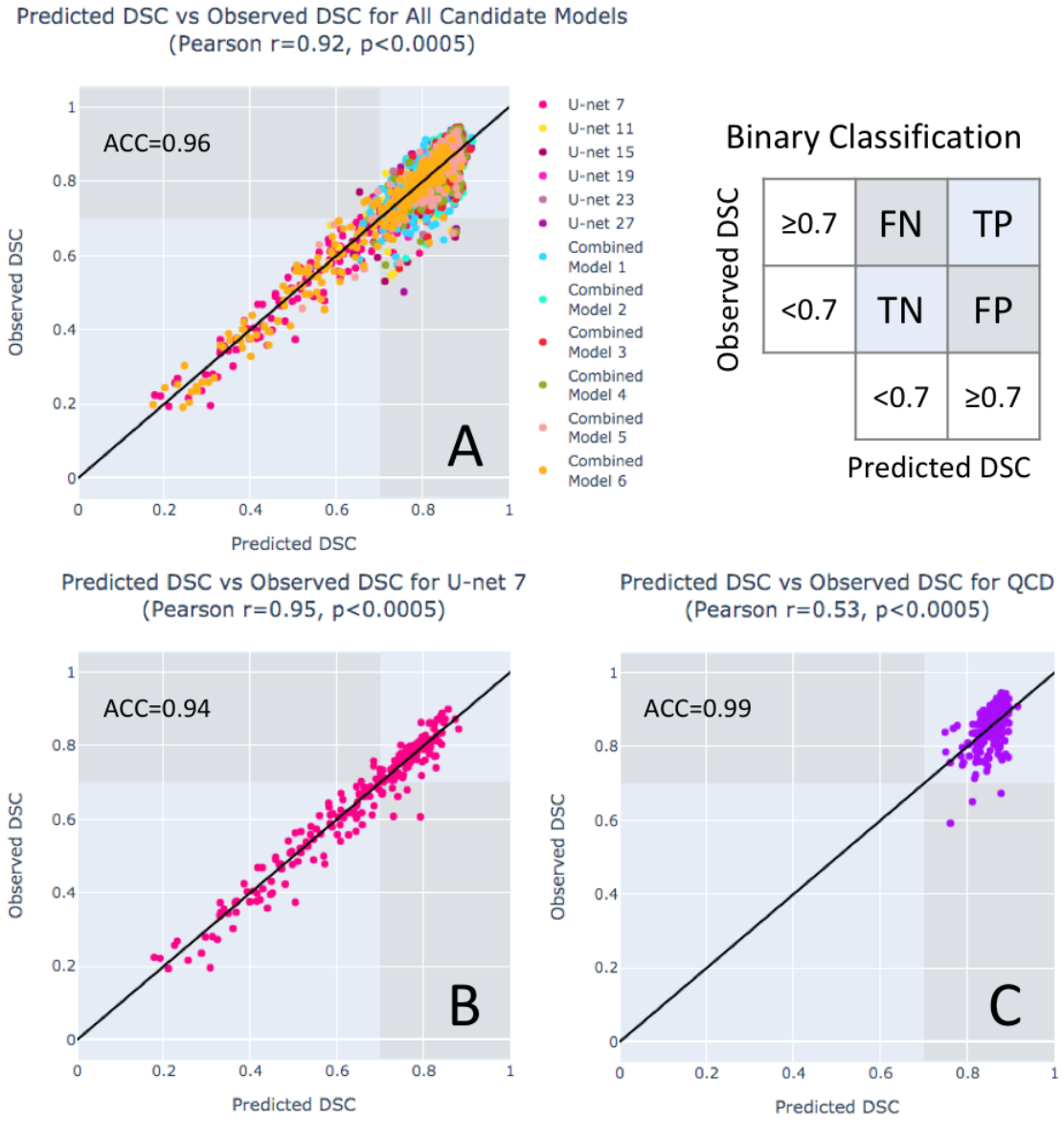


Figure 3.6: **The Pearson correlation coefficient is not necessarily an accurate indicator of quality prediction performance.** Scatter plots of the predicted vs the observed DSCs are shown for: (A) all the candidate segmentations in the QCD framework, (B) U-net 7, and (C) the final segmentations selected by the QCD framework. The highest classification accuracy (ACC) of good (observed DSC ≥ 0.7) and bad (observed DSC < 0.7) segmentations is seen in (C) the final segmentations selected by the QCD framework (ACC=0.99), compared to (A) all the candidate segmentations (ACC=0.96) and (B) U-net 7 (ACC=0.94). Although high correlations were observed for (A) all the candidate segmentations ($r = 0.92$) and (B) U-net 7 ($r = 0.94$), a much weaker correlation was obtained for (C) the final QCD segmentations ($r = 0.53$), which had a better segmentation performance (observed DSC between 0.59-0.95) and despite having the highest accuracy (ACC=0.99).

3.3.3 Accuracy of Segmentation Quality Control

The MAEs for the DSC prediction ranged from 0.031 to 0.046, for all implemented models (Table 3.2), indicating overall good prediction of quality control for all the candidate segmentations, substantiating the validity of the QCD framework. The MAE in predicting the DSCs for the QCD framework was 0.034 (Table 3.2).

The Pearson correlation of the predicted DSCs and the observed DSCs was calculated for each model (Table 3.2), and is often used to assess the performance of segmentation quality control methods. Figure 3.6A shows high correlation ($r = 0.92, p < 0.0005$) for the DSC prediction of all the candidate segmentations. This indicates that the DSC prediction can estimate a wide range of segmentation quality for all the candidate segmentations. Interestingly, the correlations measured individually for the segmentation models (Table 3.2) show that the Pearson correlation tended to be stronger if the segmentation model performed worse in terms of mean DSC, and, conversely, weaker if the segmentation model performed better. Figures 3.6B and C explain the relation using the scatter plots of the predicted DSCs and the observed DSCs for U-net 7 and the QCD final segmentations, respectively. For the shallowest U-net 7, a strong linear correlation ($r = 0.95, p < 0.0005$) can be clearly observed as the data points spread along the identity line from 0.19 to 0.90 (Figure 3.6B). However, for the QCD final segmentations, the Pearson correlation ($r = 0.53, p < 0.0005$) of quality control was weak (Table 3.2) despite the high mean DSC and the low MAE in the DSC prediction, as the data points in the scatter plot cluster around 0.59 to 0.95 (Figure 3.6C). Therefore, the Pearson correlation is not necessarily a good metric for evaluating the quality control component in this work, and may be misleading when the accuracy of the segmentation models is very high.

The DSC prediction in the QCD framework was further evaluated for binary classification of good (observed DSC ≥ 0.7) and bad (observed DSC < 0.7) segmentations. The evaluation showed high classification accuracy (ACC) for all the candidate segmentations (ACC=0.96, Figure 3.6A), U-net 7 (ACC=0.94, Figure 3.6B), and the final segmentations selected by the QCD framework (ACC=0.99, Figure 3.6C). High

Table 3.3: Agreement of estimated T1 values using the automated QCD segmentation compared with manual segmentation in the testing data. MAE: mean absolute error

Pearson Correlation	**Mean Error (SD)**	**MAE (SD)**
0.987 ($p < 0.0005$)	-4.6ms (16.7)	11.3ms (13.0)

true positive rates (TPR) were also achieved: 0.99 for all the candidate segmentations, 0.94 for U-net 7, and 1.00 for the QCD final segmentations. In addition, the false positive rates (FPR) were reported: 0.25 for all the candidate segmentations, and 0.04 for U-net 7. Only 3 false positive cases, with high predicted DSCs (≥ 0.7) but low observed DSCs (< 0.7), were found for the 220 QCD final segmentations. These results demonstrated that the DSC prediction can differentiate good and poor segmentations for quality control purpose.

The 3 false positive cases for the QCD segmentations were identified (Figure 3.7). The automatic segmentations (Figure 3.7A-C) for these cases appeared acceptable after review for practical use despite having low observed DSCs. The manual segmentation masks (Figure 3.7D-F) were excessively thin, potentially due to attempts by the human operator to avoid partial volume when myocardial coverage was not considered critical [65, 24]. This contributed to the low observed DSCs due to little overlap between the automatic segmentations and the thin manual masks. Despite the low DSCs, the myocardial T1 values estimated by the QCD agreed with the manual estimation to within ±6.5%.

3.3.4 T1 Value Estimation

The QCD achieved the highest mean DSC (Table 3.2), and thus was chosen for estimating the LV myocardium T1 values in the testing data. The result showed a high degree of agreement for the estimated T1 values between manual and automatic segmentations, with a mean error of -4.6ms, a mean absolute error (MAE) of 11.3ms, and a Pearson correlation $r = 0.987$ ($p < 0.0005$, Table 3.3).

The Bland-Altman plot (Figure 3.8) showed consistent estimation of the T1 values, with a 95% confidence interval (CI) from -3.58% to 2.72% for the differences between

the automatic and the manual segmentations. There was no apparent correlation between the T1 estimation error and the average T1, indicating that the error was not dependent on the T1 value. Further investigation found 11 outlier cases outside the 95% CI range in the Bland-Altman plot (Figure 3.8), where 7 cases were classified as 'poor' image quality, and 4 were 'acceptable'.

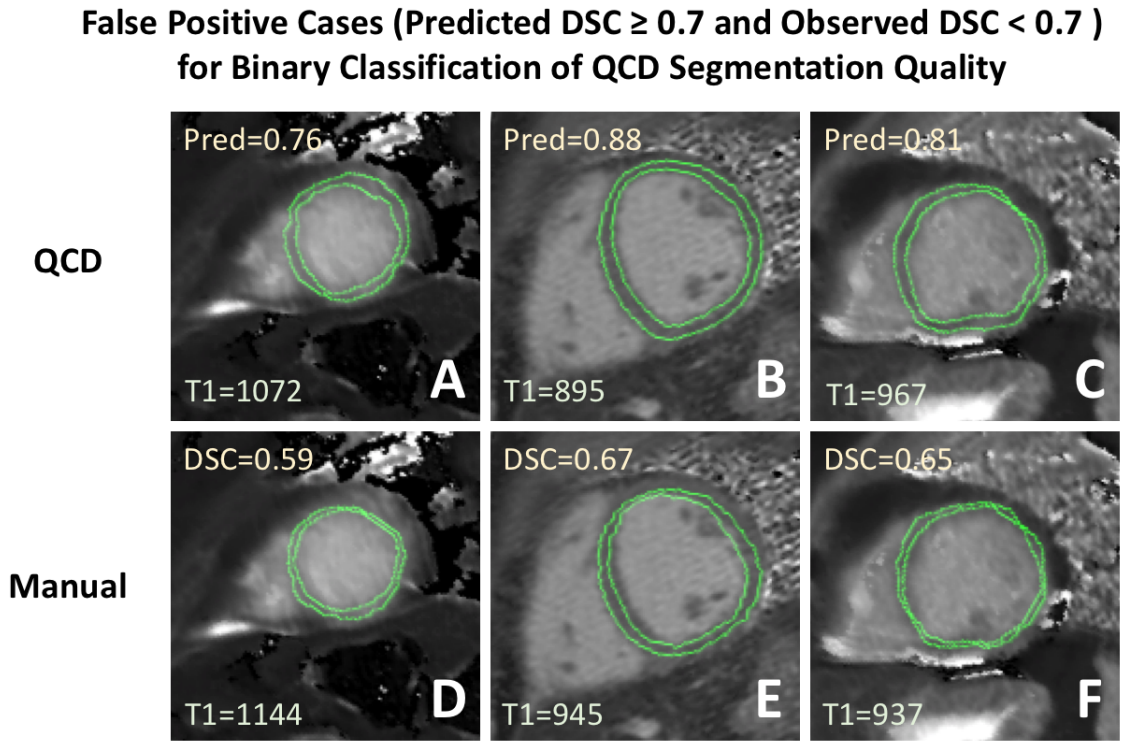


Figure 3.7: False positive cases (high predicted DSC but low observed DSC) for binary classification of the QCD final segmentation quality. The segmentations in the top row (A, B, C) were output by the QCD framework. These contours appeared acceptable and had high predicted DSCs (≥ 0.7). The bottom row (D, E, F) shows the corresponding manual contours, which appeared excessively eroded by the human operator, a valid approach in some studies aiming to limit partial volume effects. In these cases, the observed DSC values were "unfairly" low due to the low overlap between the narrow manual myocardial segmentations and the corresponding QCD outputs. Despite the low DSCs, the myocardial T1 values estimated by the QCD segmentations agreed with the manual estimations to within $\pm 6.5\%$.

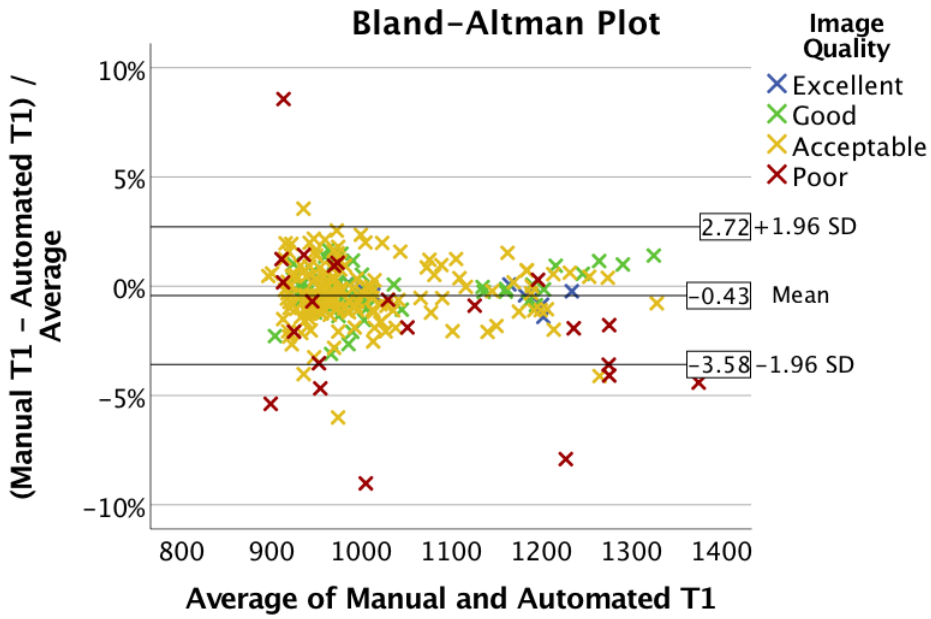


Figure 3.8: Bland-Altman plot of agreement between T1 values estimated using automated and manual segmentations. Different colours indicate the image quality as perceived by the expert human operator. Most of the points were in the range of -3.58% to 2.72% difference. Cases outside of this range were "poor-quality" (7 cases) and "acceptable" (4 cases).

3.4 Discussion

The novel real-time QCD approach was successfully applied to CMR T1 mapping automated image segmentation, with speed, accuracy, reliability, and visualisation for the purposes of real-world diagnostic medical imaging. This was demonstrated by the use of the per-case DSC prediction to select the most optimal segmentation on-the-fly from multiple intermediate candidates. This QCD framework achieved high agreement in myocardial T1 values between the automated and the manual segmentations. Furthermore, the fast processing speed of 0.39s/image enables real-time clinical applications. In addition, the analysis of the Pearson correlation and the segmentation performance exposed an undesirable dependence between the two, showing that the Pearson correlation may not always be a suitable evaluation metric for quality prediction.

3.4.1 Comparisons with Related Work

The QCD framework demonstrated high accuracy in estimating LV myocardial mean T1 value in CMR images. This framework showed high consistency with manual estimation of the myocardial T1 value, compared to the inter-observer variability between two human operators using the same T1-mapping method in [64], which reported a Pearson correlation of 0.92 with the 95% CI of relative errors ranging from -4.7% to 3.3%. The QCD framework also showed a higher Pearson correlation at 0.99 ($p < 0.0005$) for T1 estimation than that reported by [18] ($r = 0.72, p < 0.0001$) and by [43] ($r = 0.98, p < 0.01$). [60] reported a small error in estimating T1 values, with the mean relative absolute error of 4.6%. However, only 10 healthy subjects were studied in their work, which may not reflect the adaptability of their method to the real-world clinical setting where a wide range of pathologies exist. The QCD framework showed a small mean absolute error (11.3 ms) in estimating myocardial T1, even smaller than that reported by [43] for the U-net implementation (75.75 ms), and for the probabilistic network implementation (12.40 ms).

The QCD framework demonstrated high accuracy in quality control by predicting the DSC of the segmentation, regardless of the availability of manual segmentation as the ground truth. We identified existing CMR image segmentation quality control frameworks for comparison, though it is important to note that the training and testing data used were different. We achieved low MAE (0.0339) compared with the RCA quality control frameworks [31, 32, 33], in which the reported prediction MAE was at least 0.044. A CNN-regression approach [35] also reported a higher MAE (0.055) in predicting the segmentation of the LV myocardium. The low MAE of the DSC prediction achieved by the QCD framework also compares favorably with the dropout-based quality control method [37], which appeared to have a high discrepancy in predicting DSC. Unlike the QCD framework, the dropout-based approach does not have the advantage to utilise regression for more accurate DSC prediction, besides the randomness of output inherent to this approach.

The binary classification of good (observed DSC ≥ 0.7) and poor (observed DSC < 0.7) segmentations demonstrated a high classification accuracy of 0.96 for all the candidate segmentations and 0.99 for the final segmentations in the QCD framework. This is on par with the results (classification accuracy of over 0.95) reported by [33].

The whole framework (both the segmentation and the quality control) is faster (0.39 s/image) than the RCA framework, which required 11 minutes to process a single image [33]. Expectedly, the QCD framework, which utilised 6 fully convolutional neural networks, was slower than the single CNN used in [35], but only by a small fraction of a second. This demonstrates that the fast processing speed of the QCD framework permits real-time clinical applications.

3.4.2 Limitations and Future Work

The single final segmentation selection mechanism in the QCD framework is flexible to include different segmentation methods, and techniques to combine segmentations. Further research can be done to assess potential benefits of incorporating a more diverse variety of segmentation methods such as active contour models [55], or multi-atlas segmentation [79]. The use of different segmentation algorithms can potentially

further strengthen the reliability of the segmentation and the quality control of the framework by imposing anatomical constraints used in active contour models or multi-atlas segmentation. Furthermore, future research can investigate the inclusion of different techniques to combine single model segmentations, such as by weighted averaging, as candidates to be chosen as the final output in the QCD framework. With ever-advancing research in medical image analysis, one of the strongest points of this framework is that it can incorporate any prior and future classification models as intermediate solutions, which may further improve both accuracy and reliability of the overall classification process. In addition, research on better selection and choice of candidate segmentation algorithms can be beneficial in further optimisation of the QCD framework.

In this work, I focused on the quality control of automated segmentation, as a first step towards clinical translation of automated image post-processing. In the future, we aim to adapt the presented quality control-driven framework to ensure reliability of the extraction of clinical parameters from multimodal data.

The performance comparisons of the segmentation and the quality control methods between various publications need to be treated with care due to potentially significant differences of the datasets. The work presented is a proof-of-principle of the QCD framework, derived using internal datasets; further training and validation, including head-to-head comparisons of segmentation and quality control performance, using large-scale external datasets, such as the UK Biobank, will be beneficial for wider generalisability, and is future work in the pipeline.

3.4.3 Clinical Impact

Assuming equal variation in the automatic and the manual estimates, the reported 95% CI range here in the Bland-Altman plot (Figure 3.8) is narrower than the 95% CI (-4.7% to 3.3%) between two human operators processing another T1 mapping dataset in [64]. The Bland-Altman plot shows a bias of -0.43% (mean error = -4.6 ms; Table 3.3), smaller in magnitude than the reported -0.7% in [64]. Thus, the

automation can outperform the manual T1 estimation in terms of interobserver variability. In addition, the mean absolute error of 11.3 ms (Table 3.3) in T1 estimation was small, within 1 standard deviation from the figure (8.4 ± 6.3 ms) reported in [24] between two experienced operators processing only good-quality mid-ventricular T1 maps. With such a small discrepancy, the automation can be deployed reliably to extract myocardial T1 for clinical research studies. Furthermore, the automatic T1 estimation obtained a high Pearson correlation coefficient of 0.99 (Table 3.3), higher than the agreement (Pearson $r = 0.92$) between two human operators [64]. Such high agreement in estimated T1 values between the automated and the manual segmentations implies that the automated segmentation can minimise the burden of manual processing and improve time efficiency in both real-time clinical practice and large-scale research, to consistently extract T1-related clinical parameters at the level of human operators. For real-time clinical application, the framework could be integrated into MRI scanners to generate an immediate segmentation after an image is acquired for instant availability for interpretation. For large-scale clinical research and trials, the automation of segmentation can reliably process tens of thousands of datasets, saving labour-intensive processing and costs for processing large-scale imaging databases.

Across all these applications, there is an added benefit from the highly accurate quality prediction, which can reduce the effort to manually screen the data for any suboptimal results. Future work is pending to establish relevant quality thresholds, to further improve reliability of the automated segmentation to identify error-prone datasets in large-scale clinical data. This will help improve robustness to detect and interpret outlier data without excessive workload on human observers to manually score data quality. With improved quality of clinical parameters and reduction in errors, it may reduce sample sizes required for expensive clinical studies or trials, saving resources.

3.4.4 Conclusion

The QCD framework for automated quality prediction improves the accuracy and the robustness of the segmentation. The quality control exploits differences among models to predict each segmentation quality, without the need for manual contour ground truth. The predicted quality score can also be used for binary classification of segmentation quality. The selection of the most optimal segmentation is performed on-the-fly using the quality prediction, and significantly improves the accuracy above any individual network or their combinations. The proposed segmentation agreement visualisation provides a simple tool to monitor the quality control process. The validation on the CMR T1 mapping data shows wider adaptability of the framework. The automated estimates of T1 relaxation times showed near-perfect agreement ($r = 0.987, p < 0.0005$; mean absolute error (MAE) of 11.3ms) with the manual estimation used in clinical research, with a fast processing speed of 0.39s/image. The use of the QCD framework could lead to real-time parameter extraction in clinical practice and automation of labour-intensive tasks in large-scale clinical research and trials. This can enable clinicians and healthcare personnel to spend more time with patients rather than performing tedious segmentation and quality control tasks.

Candidate's Contribution

Conceptualisation, Methodology, Software, Experiments, Analysis, Data curation, Literature reviews, Writing - original draft, Writing - review and editing

Publications

1. **Hann, E.**, Popescu, I.A., Zhang, Q., Gonzales, R.A., Barutçu, A., Neubauer, S., Ferreira, V.M., Piechnik, S.K., 2021. Deep Neural Network Ensemble for On-the-Fly Quality Control-Driven Segmentation of Cardiac MRI T1 Mapping. Medical Image Analysis. 102029. https://doi.org/10.1016/j.media.2021.102029

2. **Hann, E.**, Popescu, I.A., Zhang, Q., Barutçu, A., Neubauer, S., Ferreira, V.M., Piechnik, S.K.. Quality Control-Driven Artificial Intelligence for Reliable Automatic Segmentation of T1 Mapping Images, in: SCMR/ISMRM Co-Provided Workshop, Orlando, Florida, United States. 12 - 15 February 2020. Published: 12 February 2020.

3. **Hann, E.**, Piechnik, S.K., Popescu, I.A., Zhang, Q., Werys, K., Ferreira, V.M. "Method and Apparatus for Quality Prediction". Patent application submitted to Oxford University Innovation (project 16045) PCT/GB2020/050249 , Filed 4 February 2020

Chapter 4

Generalisability of the QCD Framework to Large-Scale External Data for CMR T1 Mapping

4.1 Introduction

In Chapter 3, I demonstrated that the QCD segmentation framework can robustly delineate the LV myocardium in T1 mapping images available internally in OCMR. The QCD framework also provides accurate scalar-value prediction of the Dice similarity coefficient (DSC) to indicate the expected accuracy of segmentation on a per-case basis. For each given T1 mapping input, the QCD framework selects the optimal final segmentation from multiple samples on-the-fly based on the DSC prediction. The same DSC prediction was further implemented to perform a binary classification of the segmentation quality based on a pre-defined thresholding scheme. Segmentations which obtain predicted DSCs above a certain threshold are automatically deemed as reliable, while those below the threshold may require manual inspection and lead to further actions. This mechanism is intended to serve as a clear actionable indicator of whether a given segmentation can be used reliably for research and diagnostic purposes, under varied levels of human attention.

However, there was no guarantee of generalisation when applying to unseen external datasets, as the QCD framework had only been trained, validated, and tested using the internal OCMR T1 mapping data in Chapter 3. External datasets may have been acquired with different MRI devices, using different hardware or software versions, scanning parameter variations and operators that all could impact on the image characteristics, such as planning, absolute values, noise, etc. Ground truth segmentations for these new data sets may have been performed by different observers following different guidelines. These differences can potentially undermine both the automatic segmentation and quality control performance, assessed in reference to the gold standard. Thus, it is imperative to evaluate the QCD framework on independent, unseen large-scale external datasets, such as the UK Biobank, as introduced in Chapter 1.

4.1.1 Related Work

As reviewed in Chapter 1, research on automated segmentation methods developed specifically for CMR T1 mapping has progressed rapidly in the recent years [18, 60, 43]. Two of the methods have shown promise on using neural networks to perform automated segmentation [18, 43]. Further, segmentation quality control mechanism can be implemented upon probabilistic networks [42] by using an additional convolutional neural network, based on [80], for quality classification, trained using manual annotations of segmentation correctness [43]. However, the current state-of-the-art methods have only used the same datasets randomly split for both development and testing, thus fall short on capturing the performance under challenges arising from generalising to other independent large-scale datasets.

Beyond CMR T1 mapping applications, there has been increasing research interest in generalisation of uncertainty estimation methods, which can be used to quality control automated segmentations for cardiac T1 mapping. A simple evaluation of generalisation for uncertainty estimation could involve testing the method using unseen datasets from a different domain, for example, images acquired at a different centre [38, 21]. Another way to evaluate generalisation is to test the method using unseen datasets generated with different extent of distortion [81], or even out-of-distribution testing datasets [81, 21, 82, 46]. Compared with other uncertainty estimation methods, such as Monte Carlo dropout-based Bayesian neural networks [38], deep ensembles [45] have been shown to have better generalisation performance [81, 21, 82]. The additional capacity of deep ensembles to cope with broader ranges of data has been apportioned to the higher diversity in generated predictions [46].

4.1.2 Objectives

In this chapter, the QCD framework, trained using the internal OCMR ShMOLLI native T1 mapping data in Chapter 3, is evaluated using the external UKBB dataset as a test of generalisation for both segmentation and quality control components.

4.2 Material and Methods

4.2.1 Testing Dataset: The UK Biobank T1 Mapping Data

As a part of the UKBB Imaging Component, a total of 2020 ShMOLLI native T1 maps were available. The UKBB T1 maps have been confirmed good image quality (no more than 1 segment affected by image artefacts) and manual segmentations annotated by a single image analyst (EL). The observer has more than 15 years of CMR image analysis experience. The image analyst has not contributed to the ground truth in development dataset for the QCD framework, which was acquired and annotated by multiple observers in OCMR until 2017. Thus, the new material is wholly independent of the validation and testing dataset.

The UKBB material was acquired with different scanning parameters ("Work In Progress" (WIP) package 780B) at 1.5T (MRI Scanner: Siemens Aera, Germany) from the OCMR material (WIP 561, 448C, and 1024B) at both 1.5T and 3T, potentially resulting in differences in T1 estimates [61, 4, 43]. The UKBB T1 maps were acquired with a flip angle of 35 degrees, TR = 2.6 ms, TE = 1.07 ms, GRAPPA factor = 2, interpolated voxel size of $0.9 \times 0.9 \times 0.8$ mm^3, and a typical field of view = 360×236. Besides the acquisition parameters, the UKBB focused on population study with mostly healthy subjects while the OCMR focused on clinical scans with more pathological data, potentially impacting myocardial T1 range and morphology of the myocardium. These differences between the training material and the evaluation material may typically impact the performance of the automatic segmentation. Thus, the UKBB material can test the robustness of the QCD framework trained on the OCMR data.

4.2.2 Trained QCD Framework for Segmentation and Quality Prediction

The QCD framework (details in Chapter 3) to be evaluated in this chapter has already been trained on a total of 1906, manually-contoured, ShMOLLI native T1 maps available internally in OCMR. Similar to Chapter 3, the QCD framework and candidate

models are evaluated with the ground truth contours to calculate observed DSCs of the segmentations. Quartiles, lower fences (lower quartile – 1.5 times interquartile range), and box plots of the observed DSCs are also provided to aid the evaluation. The observed DSCs are subsequently used as the ground truth to evaluate the DSC prediction.

4.2.3 Automated Segmentation Quality Classification

Based on the ground truth segmentation quality labels established, an approach similar to [31] was employed. To establish the ground truth segmentation quality labels, the segmentations were classified as follows:

- **Good quality:** if the observed DSC ≥ 0.8;
- **Minor issues:** if the observed DSC ≥ 0.6 but < 0.8;
- **Major issues:** if the observed DSC < 0.6.

In comparison to the single threshold at 0.7 used in Chapter 3, this scheme provides not only a more stringent standard for "good quality" segmentations, but also flexibility to evaluate usefulness of segmentations predicted with "minor issues" as acceptable without manual review.

To detect "good quality" segmentations in the absence of the ground truth, the DSC prediction can be utilised and extended to a binary classification task, with an optimal threshold. In this chapter, I established an optimal threshold for predicted DSC using a receiver operating characteristic (ROC) curve and Youden index.

The established thresholds allow to study how segmentations of varying quality agree with the manual ground truth in estimating the LV myocardial T1. The segmentation quality classification is evaluated for the following markers of the predictive properties:

- true positive rate ($TPR = \frac{TP}{TP+FN}$)
- false positive rate ($FPR = \frac{FP}{FP+TN}$)

- accuracy ($ACC = \frac{TP+TN}{TP+FP+FN+TN}$)
- precision ($PRC = \frac{TP}{TP+FP}$)

, in which

- true positive (TP): "good quality" and predicted DSC $\geq$ the optimal threshold
- true negative (TN): "minor issues" or "major issues" and predicted DSC $<$ the optimal threshold
- false positive (FP): "minor issues" or "major issues" but predicted DSC $\geq$ the optimal threshold
- false negative (FN): "good quality" but predicted DSC $<$ the optimal threshold.

4.2.4 Automated Myocardial T1 Estimation

In this application, the clinical parameter to be extracted is the LV myocardial T1 relaxation time. Therefore, assessing the quality of automated T1 estimations is of crucial importance in extracting useful clinical parameters. I performed this evaluation by comparing the T1 value extracted using the QCD segmentation (QCD T1) with the ground truth T1 value extracted using the manual segmentation (manual T1). This is assessed using a Bland-Altman plot where the x-axis is the ground truth T1 and the y-axis is (QCD T1 – manual T1) divided by the average of QCD T1 and manual T1. There are separate Bland-Altman plots for cases with predicted DSC equal or greater than the optimal threshold and cases with predicted DSC below the optimal threshold.

4.3 Results

4.3.1 Segmentation and Quality Prediction

Table 4.1: Segmentation and DSC prediction performance for the QCD and candidate models.

Segmentation Model	Mean DSC (SD)	Median (LQ - UQ)	MAE
U-net 7	0.733 (0.084)	0.747 (0.690 - 0.792)	0.026
U-net 11	0.827 (0.055)	0.839 (0.802 - 0.866)	0.029
U-net 15	0.838 (0.050)	0.846 (0.815 - 0.870)	0.030
U-net 19	0.845 (0.055)	0.855 (0.827 - 0.877)	0.032
U-net 23	0.835 (0.065)	0.847 (0.817 - 0.872)	0.035
U-net 27	0.832 (0.074)	0.846 (0.816 - 0.871)	0.033
Combined Model 1	0.774 (0.071)	0.787 (0.747 - 0.817)	0.042
Combined Model 2	0.832 (0.047)	0.838 (0.809 - 0.864)	0.037
Combined Model 3	0.846 (0.044)	0.852 (0.826 - 0.876)	0.032
Combined Model 4	0.854 (0.050)	0.863 (0.836 - 0.884)	0.028
Combined Model 5	0.848 (0.063)	0.863 (0.830 - 0.885)	0.027
Combined Model 6	0.763 (0.099)	0.784 (0.723 - 0.829)	0.036
QCD	0.848 (0.044)	0.854 (0.827 - 0.877)	0.032

Remarks: SD = standard deviation, LQ = lower quartile, UQ = upper quartile, MAE = mean absolute error

The performance for segmentation and quality control is reported in Table 4.1. The QCD framework (mean DSC=0.848; median DSC=0.854) obtained good overall segmentation performance following Combined Model 4 (mean DSC=0.854; median DSC=0.863) and Combined Model 5 (mean DSC=0.848; median DSC=0.863). Although Combined Model 3 had the best overall segmentation performance in Chapter 3, it was outperformed by the QCD segmentation in terms of median DSC (sign test, $p < 0.0005$) on the UKBB data. Similarly for the single U-net models, U-net 19 (mean DSC=0.845; median DSC=0.855) obtained the highest mean DSC and median DSC, outperforming deeper U-net 23 (mean DSC=0.835; median DSC=0.847) and U-net 27 (mean DSC=0.832; median DSC=0.846), which had better performance on the validation data in Chapter 3. These demonstrate the change in performance profiles among candidate models when applying the framework to a new dataset. For evaluation of quality prediction, mean absolute error (MAE) is reported for each

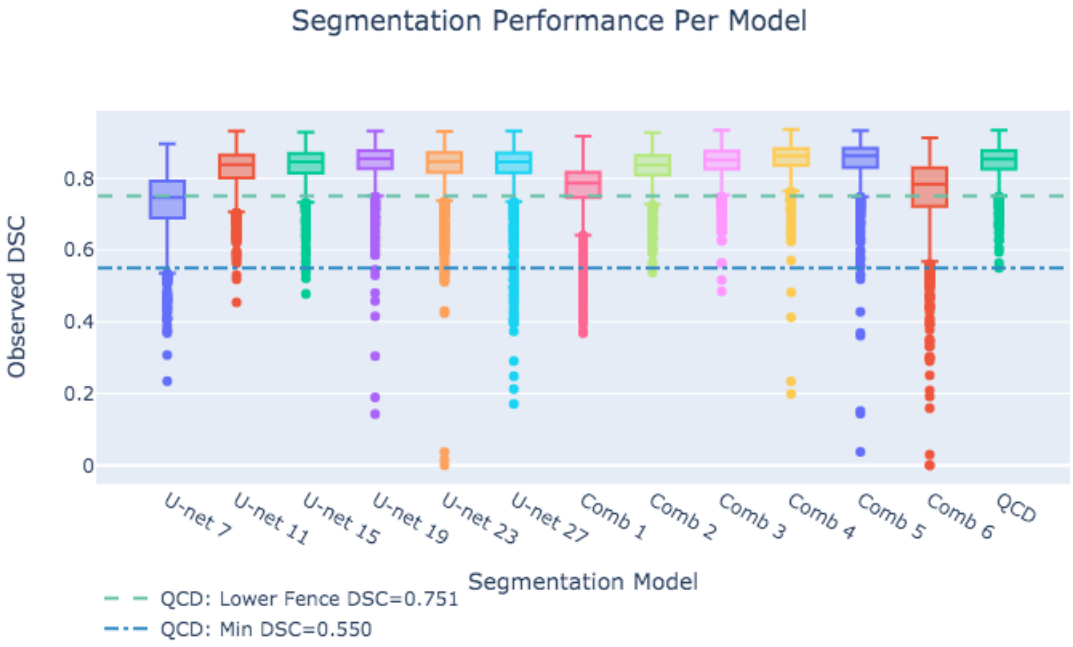


Figure 4.1: Segmentation performance is shown for each segmentation model. The QCD segmentation achieved robust DSC performance with one of the highest for the lower fence (green line) and the highest for the minimum observed DSC (blue line).

model comparing predicted DSC and observed DSC. All reported mean absolute errors were within 0.042, indicating accurate DSC prediction for all components of the QCD framework.

Figure 4.1 shows a box plot for each candidate model and the QCD framework. It was demonstrated that the QCD framework had higher segmentation robustness, relative to other segmentation candidates. The QCD framework achieved one of the highest values for the lower fence of observed DSC. Further, it achieved the highest value for the minimum observed DSC at 0.550, demonstrating excellent segmentation robustness.

4.3.2 Segmentation Quality Classification Using DSC Prediction

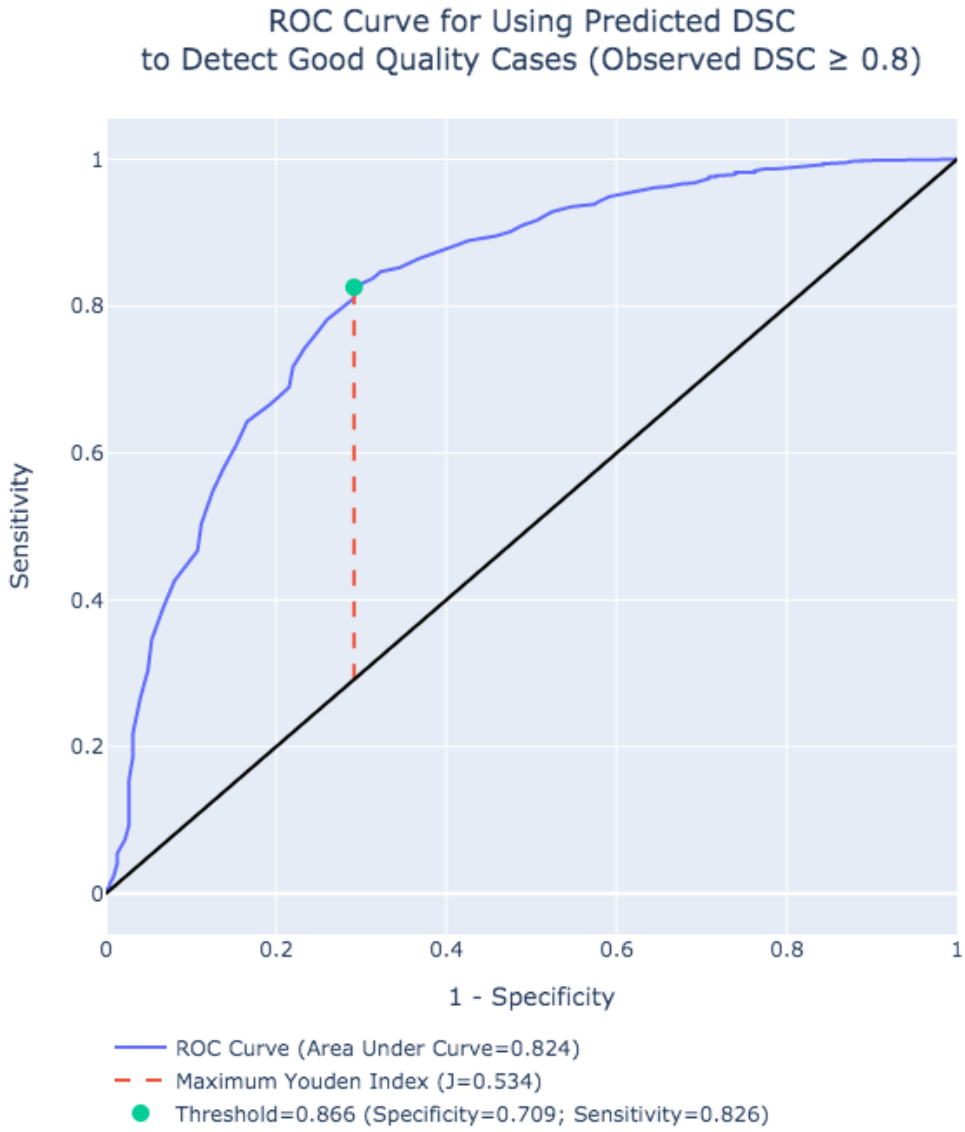
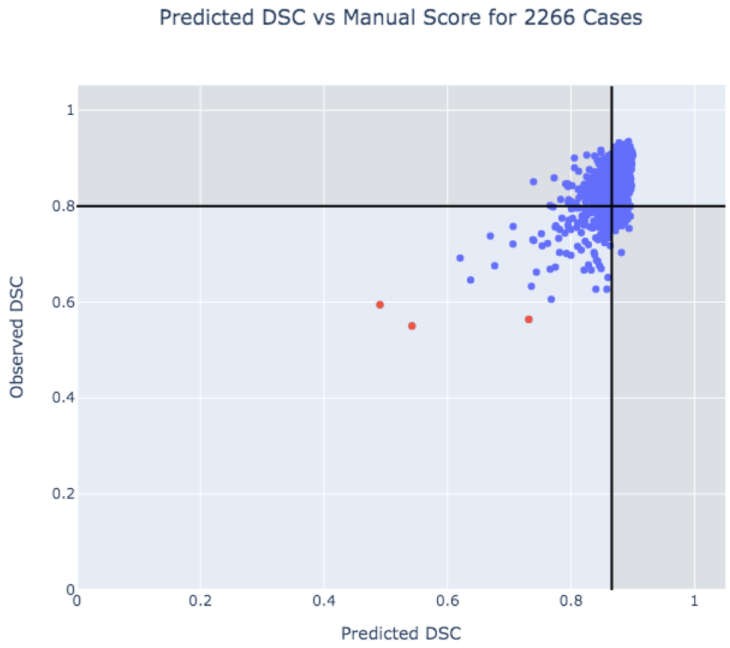


Figure 4.2: Receiver operating characteristic (ROC) for predicted DSC ability to detect good quality cases (observed DSC $\geq$ 0.8). Area under the curve (AUC) is 0.824. The maximum Youden index (J=0.534, red dash line, corresponding to predicted DSC threshold of 0.866) indicates an optimal exchange between specificity of 0.709 and sensitivity of 0.826.



		Predicted DSC	
		< 0.866	≥ 0.866
Observed DSC	≥0.8	FN: 313 (15.5%)	TP: 1484 (73.5%)
	<0.8 & ≥0.6	TN: 155 (7.7%)	FP: 65 (3.2%)
	<0.6	TN: 3 (0.1%)	FP: 0 (0%)

Figure 4.3: A scatter plot (left) of predicted DSC (x-axis) and observed DSC (y-axis) for the QCD framework tested in the independent material of 2020 UK Biobank T1 maps. 3 segmentations with "major issues" are marked as red. A horizontal black line and a vertical black line respectively denote the observed DSC threshold (0.8) and the predicted DSC threshold (0.866) for segmentation quality classification. The table (right) summarises the statistics of the quality classification results.

The ROC curve (Figure 4.2) shows the performance of the DSC prediction in the QCD framework to detect "good quality" segmentations (observed DSC$\geq$ 0.8) in the UK Biobank material. The area under the curve (AUC) is 0.824 ($p < 0.0005$), indicating good performance for the detection task. The maximum Youden index (J=0.534) is shown as the red dash line in Figure 4.2 to indicate the optimal predicted DSC threshold at 0.866 (the green dot in Figure 4.2) for detecting "good quality" segmentations, with sensitivity of 82.6% and specificity of 70.9%.

The optimal threshold of 0.866 on predicted DSC was subsequently used to perform automatic binary classification of segmentation quality to detect "good quality" segmentations (observed DSC $\geq$ 0.8). A scatter plot (Figure 4.3 left) shows predicted DSC as the x-axis and observed DSC as the y-axis, with a vertical black line and a horizontal black line respectively denoting the predicted DSC threshold (0.866) and

the observed DSC threshold (0.8). The binary classification results are shown in the table (Figure 4.3 right). There were 1484 true positive cases, accounting for the majority (73.5%) of all segmentations. There were only 65 false positive cases, accounting for 3.2% of all segmentations. None of the false positive cases had "major issues" (i.e. observed DSC < 0.6), thus all false positive segmentations obtained observed DSCs above 0.6. False negative cases accounted for 15.5% (313 cases) of all segmentations. There were 158 true negative cases (7.8%), of which 155 segmentations (7.7%) were "minor issues" and only 3 (0.1%) had "major issues". Overall, the majority (76.7%) of the QCD segmentations were predicted "good quality", while less than 25% of them may require visual inspection by expert image analysts.

The segmentation quality classification was evaluated using classification metrics. The true positive rate (TPR) shows that 82.6% of the good quality cases were correctly detected. False positive rate (FPR) shows that 29.1% of the minor or major issues cases obtained predicted DSC ≥ 0.866, thus being falsely classified as positive by the automatic quality control. An accuracy (ACC) of 81.3% indicates that the majority of cases were correctly classified. For segmentations being classified as positive (predicted DSC ≥ 0.866), the precision of 95.8% shows that almost all these were actually good quality, having observed DSC ≥ 0.8. Thus, the QCD framework was successful in detecting truly good quality segmentations.

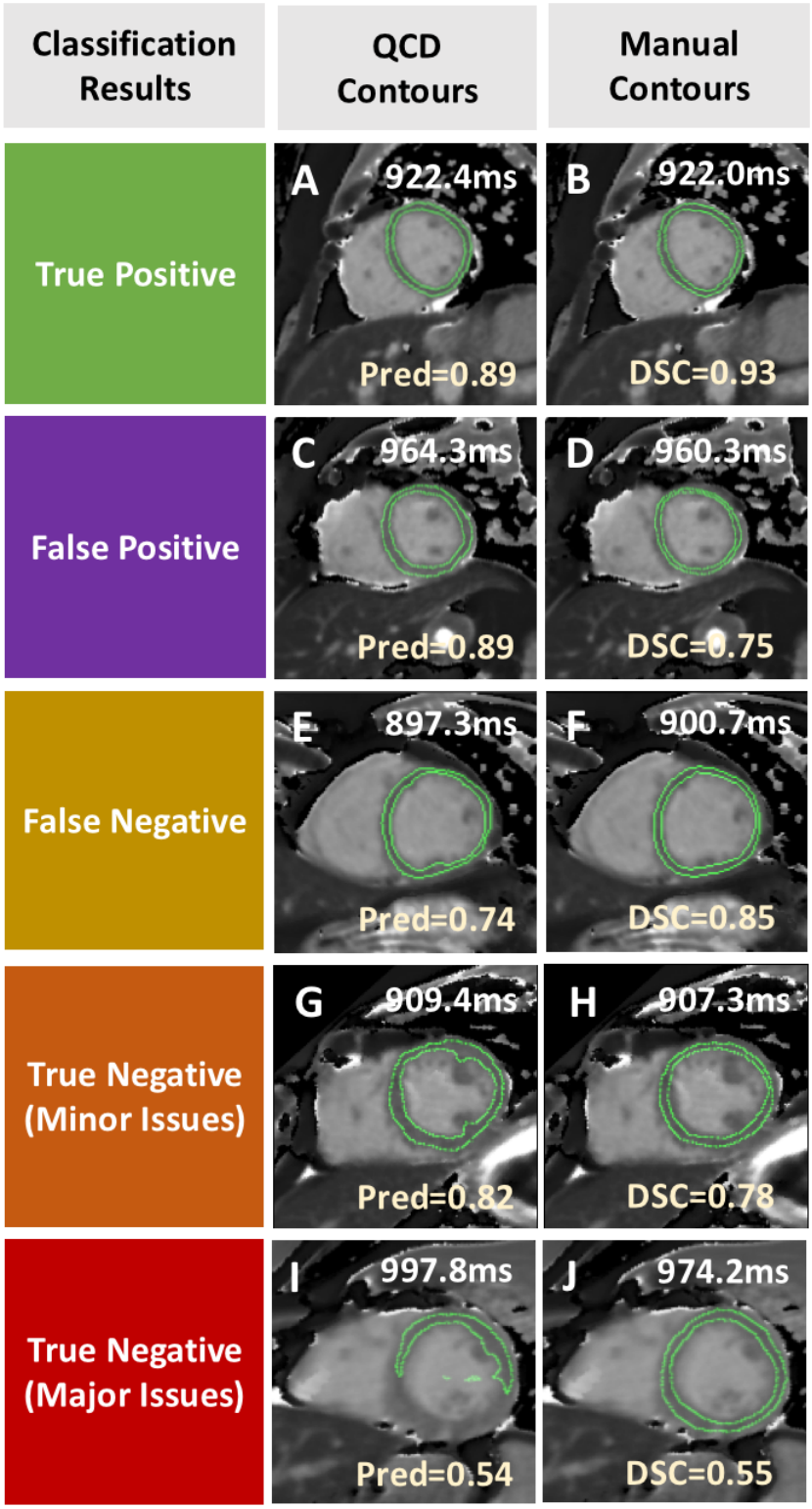


Figure 4.4: Examples for segmentation quality classification. 5 cases are shown having T1 maps overlaid with automatic QCD contours (A-E) and manual contours (F-J). The mean myocardial T1 value is also shown for each segmentation.

Five selected segmentation examples are shown in Figure 4.4 to illustrate different classification categories in the scatter plot in Fig. 4.3. Fig. 4.4A is a true positive QCD segmentation with a high predicted DSC of 0.89 and a high observed DSC of 0.93. Not only did the segmentation highly resemble the manually annotated mask (Figure 4.4B), but the average myocardial T1 value estimated (922.4 ms) also agreed closely with the manual ground truth (922.0 ms).

Figure 4.4C is a false positive QCD segmentation with a high predicted DSC of 0.89 but a low observed DSC of 0.75. This QCD segmentation appears acceptable after manually reviewing. Similar to the examples shown in Figure 3.7 in Chapter 3, the low observed DSC was due to the narrow manual mask (Figure 4.4D). Despite the low observed DSC, the T1 value estimated (964.3 ms) still agreed closely with the ground truth (960.3 ms).

Figure 4.4E is a false negative case, with a low predicted DSC of 0.74 but a high observed DSC of 0.85. The QCD segmentation appeared acceptable and mostly resembled the corresponding manual ground truth (Figure 4.4F), with high agreement on T1 estimation (the QCD-estimated T1 = 897.3 ms; manually estimated T1 = 900.7 ms). An extended information on this example is provided in Figure 4.5, which shows high disagreement among candidate segmentations, thus obtaining a low predicted DSC, discussed further below.

Figure 4.4G is a true negative case, with a low predicted DSC of 0.82 and a low observed DSC of 0.78 (minor issues). The QCD segmentation had noticeable differences compared to the manual ground truth Figure 4.4H, yet having very little difference in T1 estimation (QCD-estimated T1 = 909.4 ms; manually estimated T1 = 907.3 ms).

Figure 4.4I is a true negative case, with a low predicted DSC of 0.54 and a low observed DSC of 0.55 (major issues). The QCD segmentation failed to segment half of the myocardium, compared to the manual ground truth (Figure 4.4J). The T1 estimation difference (QCD-estimated T1 = 997.8ms; manually estimated T1 = 974.2ms) is relatively high compared to the previous cases. There are just three examples of such severe mis segmentation and will be discussed further below.

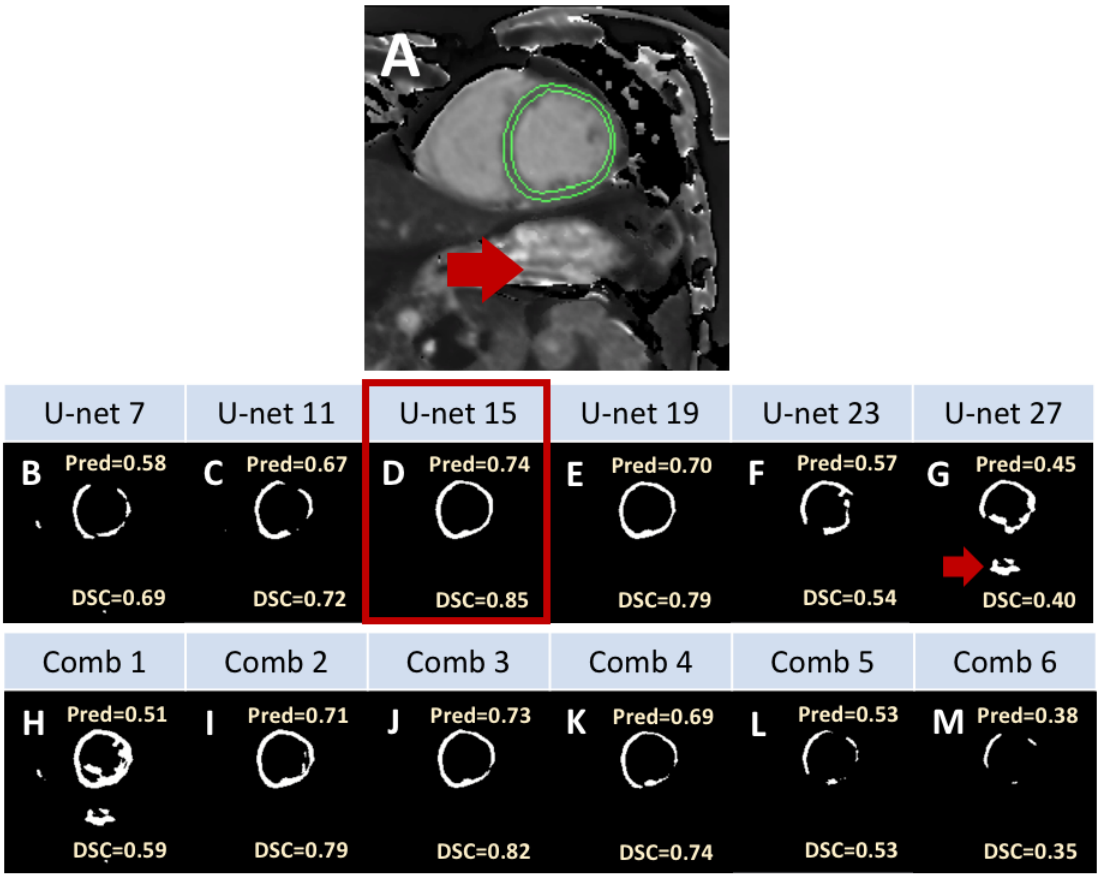


Figure 4.5: Extended example of the false negative case in Fig. 4.4C showing the manual segmentation (A), candidate segmentations (B-M) with the corresponding predicted DSCs and observed DSCs. Among all candidate segmentations, only the segmentation generated by U-net 15 (D) was able to achieve a high observed DSC of 0.85 (good quality), successfully chosen by the QCD framework as the final optimal segmentation based on the DSC prediction. However, a low predicted DSC of 0.74 was obtained, below the predicted DSC threshold of 0.866, due to high disagreement in segmentation between U-net 15 and other candidate models.

Figure 4.5 shows an extended example of Figure 4.4C to demonstrate that high disagreement among candidate segmentations that lead to underestimation in DSC prediction. Despite the U-net 15 (Figure 4.5D) having high resemblance (observed DSC=0.85) to the manual ground truth (Figure 4.5A), it was predicted as low DSC (0.74) and subsequently classified as suboptimal quality. This was potentially due to the extra-cardiac structures (stomach, as pointed by the red arrows in Figure 4.5AG) affecting the automatic segmentation. The high disagreement among the candidate segmentations led to underestimation in DSC prediction for most candidate models. Only U-net 23 (Figure 4.5F), U-net 27 (Figure 4.5G), and Combined Model 6 had overestimated predicted DSC, still within a reasonable range (≤ 0.05) which would not affect segmentation quality classification. Furthermore, what mattered was that the failures were very discordant in appearance, including Combined Model 4 and Combined Model 5, which otherwise were the best overall segmentation performers (Table 4.1). This example highlights the possibility that the best overall segmentation models can fail. This is consistent with the observation in the DSC box plots (Figure 4.1), in which Combined Model 4 and Combined Model 5 exhibited long low tails of outliers. In contrast, the QCD framework demonstrated robustness, with a higher tail of outliers in Figure 4.1. In the case of Figure 4.5, it is reassuring that the QCD chose the best option (U-net 15).

Figure 4.6 details all 3 true negative "major issues" cases shown as red data points in Figure 4.3. The QCD segmentations (Figure 4.6A-C) and manual segmentations (Figure 4.6D-F), with the corresponding predicted DSCs (< 0.866) and observed DSCs (< 0.6), are displayed. In Figure 4.6A-B, 2 QCD segmentations which failed to segment significant proportions of LV myocardium, compared to the extent identified correctly in the corresponding manual segmentations (Figure 4.6D-E). These failures are possibly due to the unusual rotation angle of the field of view of the T1 map. Another QCD segmentation (Figure 4.6C) of an apical slice is shown with the corresponding manual segmentation (Figure 4.6F). In this case, the predicted quality score seemed correct, as the QCD segmentation seemed acceptable on review. The

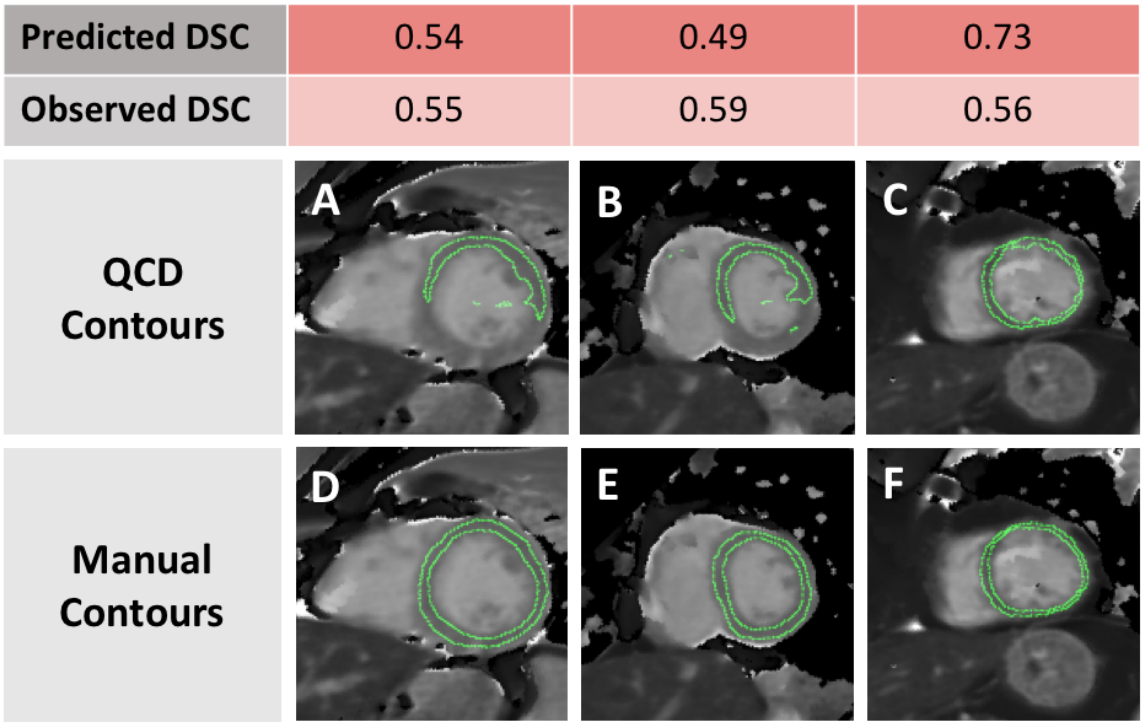


Figure 4.6: T1 maps with overlaid segmentation results for the three true negative "major issues": QCD segmentations (A-C) and manual segmentations (D-F), with the corresponding predicted DSCs (< 0.866) and observed DSCs (< 0.6).

observed DSC of 0.56 was low, which was due to the excessively thin manual segmentation mask, and this example can thus be attributed to large interobserver variability in contour placement.

To summarise, the quality prediction and classification components of the QCD framework showed very good performance. The DSC prediction achieved low mean absolute errors below 0.042 for all the segmentation models. It can be extended to quality classification with an optimal threshold at 0.866, derived using the Youden index. The quality classification detected "good quality" segmentations with high precision of over 95%, and saved over 70% of the manual effort to review the segmentations. Most importantly, the quality segmentation flagged up all "major issues" cases, which had high disagreement in the LV myocardial T1 estimation when compared with the manual ground truth.

4.3.3 Myocardial T1 Estimation

Figure 4.7 provides two Bland-Altman plots showing the average T1 value between the QCD framework and the manual ground truth (x-axis) versus T1 estimation difference (y-axis). Fig. 4.7A summarises 1549 cases with predicted DSC ≥ 0.866, both "good quality" (green) and "minor issues" (purple) cases having T1 estimation differences within $\pm 5\%$. The 95% confidence interval (CI) of the mean T1 estimation difference ranged from -0.90% to 1.93%. Fig. 4.7B shows 471 cases with predicted DSC < 0.866, where 95% CI of mean T1 estimation difference ranged from -1.75% to 4.64%. The t-test showed that the means in (A) and in (B) were significantly different ($p < 0.0005$). All "good quality" cases (blue) obtained T1 estimation differences within $\pm 5\%$, whereas the T1 differences for 18 of the "minor issues" cases (yellow) and 1 of the "major issues" cases (red) were outside the $\pm 5\%$ error relative to the ground truth. Therefore, the segmentation classification with the predicted DSC threshold of 0.866 successfully avoided cases with high disagreement in T1 estimation compared to the manual ground truth.

The distributions of T1 estimation differences are shown in Figure 4.8 for different quality classification results, with the dashed lines showing the means of T1 estimation differences. The one-way analysis of variance (ANOVA) showed that at least two of the fives means were significantly different, with $p < 0.0005$. The post hoc Tukey's test showed that only two pairs – false positive and false negative; true negative (minor issues) and true negative (major issues) – failed to reject the null hypothesis with adjusted $p > 0.05$.

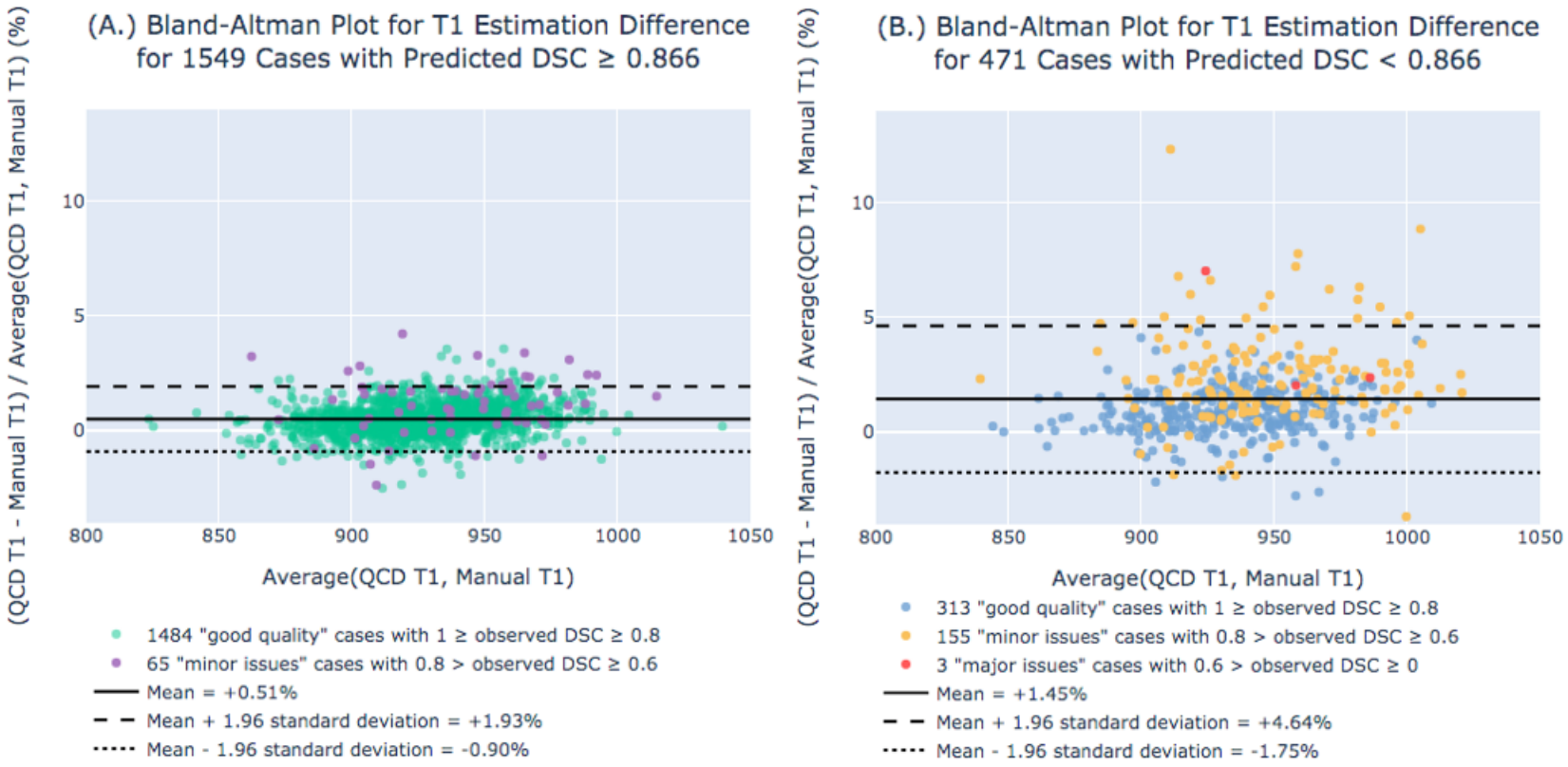


Figure 4.7: Bland-Altman plots showing ground truth T1 value (x-axis) versus T1 estimation difference (y-axis) between the QCD framework and the manual ground truth. (A) For 1549 cases with predicted DSC $\geq$ 0.866, both "good quality" (green) and "minor issues" (purple) cases were within $\pm$5% for T1 estimation difference. Approximately 95% of the cases obtained T1 estimation difference from -0.90% to 1.93%. (B) For 471 cases with predicted DSC $<$ 0.866, approximately 95% of the cases obtained T1 estimation difference from -1.75% to 4.64%. All "good quality" cases (blue) were within $\pm$5% difference, while 18 "minor issues" cases (yellow) and 1 "major issues" case (red) were outside $\pm$5% difference.

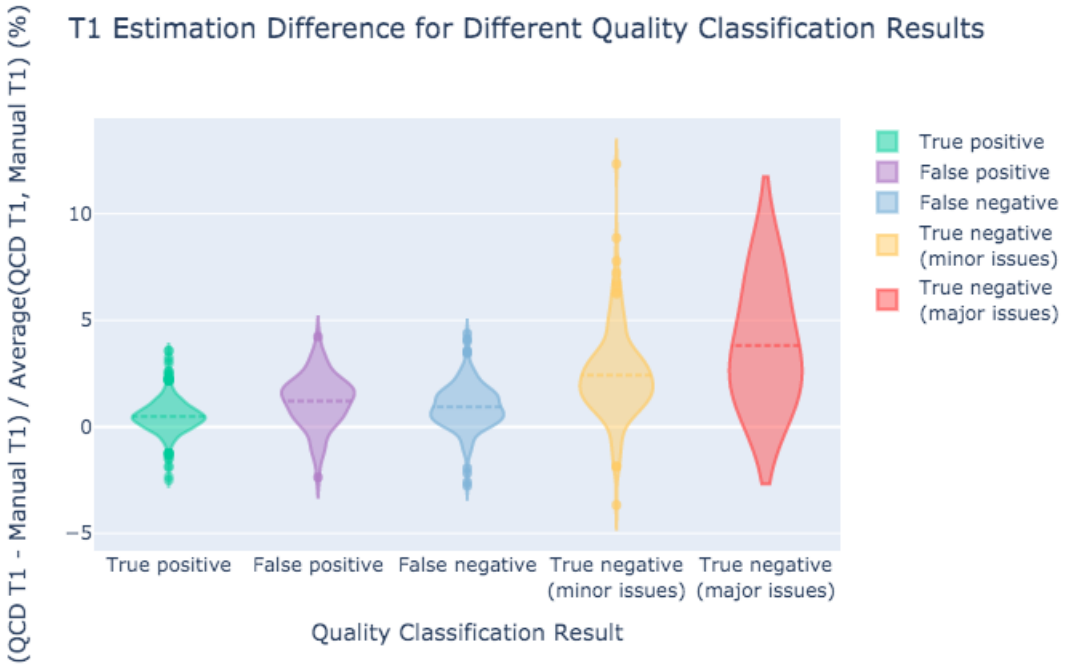


Figure 4.8: Violin plots showing the distribution of T1 estimation differences per quality classification result in the Bland-Altman plots (Figure 4.7). The dashed lines show the means of estimation differences.

4.4 Discussion

The QCD framework, trained internally on OCMR data, was successfully deployed to a large-scale external testing dataset of 2020 ShMOLLI native T1 maps, available from the UK Biobank. The QCD framework achieved excellent performance for automated segmentation, quality prediction and classification. The quality classification of "good quality" QCD segmentations achieved high precision and high recall, with an optimal threshold established on predicted Dice similarity coefficient. For QCD segmentations with high predicted DSC (≥ 0.866), the estimated myocardial T1 values were in high agreement with the manual ground truth.

4.4.1 Segmentation Performance

Compared with the performance in Chapter 3, there were noticeable differences in the ranking of the segmentation performance for the candidate models. The ranking of Combined Model 3 moved from the 1st place in Chapter 3, to the 3rd place in this chapter in terms of mean DSC among candidate segmentation models. In contrast, Combined Models 4 and 5 moved up to the top 2 in this chapter, albeit they did not make it to the top 5 in Chapter 3. These shifts in segmentation performance may be due to differences between the internal OCMR data and the external UKBB data, as described in Section 4.2.1.

The QCD framework was robust in maintaining the excellent segmentation performance. In both Chapter 3 and this chapter, the QCD segmentation achieved a high mean DSC of ≥ 0.848. Even though the QCD segmentation was outperformed by Combined Models 4 and 5 in this chapter, it was the only robust contender having achieved top 3 positions in both Chapters 3 and 4. This highlights the flexibility with which the QCD segmentation is able to maintain robust performance across varied datasets acquired using different CMR scanners at different imaging centres, and appears to be more reliable as a general approach to automatically segment previously unseen datasets.

4.4.2 Segmentation Quality Control

The QCD framework has demonstrated good performance for segmentation quality classification, with high accuracy of over 80%. For segmentations with predicted DSC of at least 0.866, excellent precision of over 95% has been achieved, implying that almost all the positive cases predicted by the QCD framework were truly good quality. Even though there were false positive cases, they only accounted for less than 4% of all the positive cases. Furthermore, it was demonstrated that false positive cases still had excellent agreement with the ground truth in estimating myocardial T1 within ±5% difference, despite the observed DSC being below 0.8. Therefore, the myocardial T1 values estimated from all the positive cases predicted by the QCD framework could be reliably used for clinical studies. As the positive cases accounted for more than 70% of all the segmentations, potentially more than 70% of the manual effort to curate good quality segmentations could be saved with the QCD framework.

4.4.3 Comparison with Related Work

Compared to existing automated processing methods of CMR T1 mapping [18, 60, 43], the QCD framework is the only method which has been thoroughly tested with a large-scale external dataset, available from the UKBB. It maintained high segmentation performance, on par with these methods [18, 43], which reported mean Dice similarity coefficient (DSC) of 0.84-0.85 for the LV myocardium segmentation. Furthermore, the QCD framework maintained excellent accuracy in estimating LV myocardial mean T1 value even for an external testing dataset. The QCD framework showed a high Pearson correlation (r) of T1 estimation, on par with r=0.98 reported in [43], higher than $r = 0.72$ in [18].

When it comes to segmentation quality prediction, the QCD framework uses a different approach than the most recent method for CMR T1 mapping [43], which employed an additional convolutional neural network to learn manual annotations of segmentation correctness. Curating such manual annotations can be time-consuming and laborious, in addition to manual contouring required to train the segmentation

component. In contrast, the QCD framework does not need any additional manual annotations of segmentation quality to train the DSC prediction for quality control. Thus, the QCD framework can be implemented in shorter time frame with lower manual labour cost compared to [43]. This is especially important when retraining is required for fine-tuning or extending to other clinical datasets, quality metrics, or applications. The QCD framework has an advantage over [43] in this regard.

With the quality classification, the segmentations with high DSCs (≥ 0.866) can provide highly accurate T1 estimation in high agreement with the manual ground truth, achieving a near-perfect Pearson correlation coefficient of 0.97, higher than the reported Pearson correlation coefficient of 0.92 between two human operators processing a different dataset of T1 maps [64]. The Bland-Altman plot (Figure 4.7A) shows a narrow 95% CI from -0.90% to 1.93% for T1 estimation difference with the manual ground truth, outperforming the human interobserver variability with 95% CI from -4.7% to 3.3% [64]. Thus, the automatic QCD framework can perform more consistent T1 estimation than human operators in terms of variability. For segmentations with low DSCs (< 0.866), the Bland-Altman plot (Figure 4.7B) shows a wider 95% CI from -1.75% to 4.64%, compared to Figure 4.7A. Therefore, the quality control component plays an important role to reduce variability and filter out potentially inaccurate results for reliable clinical parameter extraction.

4.4.4 Clinical Impact

The segmentation quality classification was successfully applied on a large-scale external dataset to detect good quality segmentations with high precision of and high recall. This helps address not only the problem of time-consuming manual segmentation, but also the problem of also time-consuming manual inspection of segmentation on a per-case basis. For this external testing dataset of 2020 T1 maps, the segmentation quality classification indicated that over 70% of the segmentations are excellent quality with no perceivable impact on the target overall T1 estimates, and thus do not require further manual inspection before use. Less than 30% of contours have been flagged for manual inspection. Currently, the QCD framework is the only automated

segmentation method for CMR T1 mapping which has successfully demonstrated robust generalisability to a large-scale external testing dataset. This serves as reassurance that the framework can scale up to larger data using compatible CMR mapping technology.

The high agreement in T1 estimation between the QCD segmentations, predicted as good quality, and the manual ground truth demonstrates that the QCD framework can be used reliably in processing clinical T1 maps with high accuracy. This is evidenced by the high Pearson correlation of 0.97, and the low mean relative error of 0.51% (SD=0.72). With improved quality of clinical parameters and reduction in errors, it may reduce sample sizes required for expensive clinical studies or trials, saving resources. In practical terms, this means large-scale datasets can be processed efficiently with reduced burden on manual processing, subject to power calculations. Extrapolating to the cohort of 100,000 cases for the UK Biobank imaging component, about 70,000 segmentations judged by the QCD to be high quality can be used immediately for most analysis, such as establishing reference ranges for healthy subjects, without the need of manual inspection. For studying rare conditions, manual contouring of 30,000 cases with potential failures in automatic segmentation may be necessary. To alleviate the burden of manually processing 30,000 T1 maps, it is possible to train an additional QCD framework specifically for these data. It is estimated that 2000 manually contoured T1 maps may be required for the training, similar to the amount of the training data used in Chapter 3. For a single observer, this may take about 2 months to complete the data curation, assuming that a single analyst takes 10 minutes to contour a single T1 map and works 40 hours a week. Once trained, the remaining 28,000 T1 maps can be automatically contoured in about 3 hours. With an assumed success rate of 70%, a total of 21,600 automatic and manual segmentations could be curated in about 2 months. Otherwise, manual contouring for the same amount of T1 maps would require at least 20 months of full-time analysis, without the additional QCD framework.

4.4.5 Limitations and Future Work

Given the flexibility of the QCD framework, prediction of different segmentation quality metrics such as DSC, Hausdorff distance can be implemented by exploiting the candidate segmentation agreement. In this thesis, only DSC prediction has been implemented for the quality control component, since it has been widely implemented in the state-of-the-art segmentation quality control methods such as [35, 38]. Implementing prediction of other segmentation quality metrics, such as Hausdorff distance, may allow further customisation for specific applications as well as improve the accuracy of quality classification.

Comparison with the state-of-the-art methods is limited as the testing datasets used are different. Even though [43] had also processed the ShMOLLI native T1 maps from the UK Biobank, it is still difficult to compare the results directly, as the cases selected for testing are likely different. It will be beneficial to have a large-scale common testing dataset available publicly for a head-to-head comparison among various state-of-the-art methods.

In this chapter, the segmentation quality classification has only been validated quantitatively against the observed DSC, which is the degree of overlap calculated between the automatic and the manual segmentations, but has not been evaluated qualitatively by a human observer. DSC prediction alone may not capture the full spectrum of criteria necessary to meet clinical requirements. For example, DSC metrics could be insensitive to a few misplaced pixels outside the structure of interest, which may not be considered acceptable in certain clinical cases. Therefore, it is beneficial to have an expert image analyst to validate the QCD framework for quality assurance. In Chapter 5, the QCD framework would be evaluated against manual scoring of segmentation quality by a human assessor.

4.4.6 Conclusion

The QCD framework for automated segmentation and quality prediction proved robust and accurate when applied to a large-scale independent unseen testing dataset

available from the UK Biobank. The results demonstrated the generalisability, flexibility and consistency of the QCD framework, compared to candidate networks in segmenting varied datasets.

I have shown that the predicted quality score can also be reliably used for binary classification of segmentation quality by finding the optimal threshold. The segmentation quality classification detects good quality segmentation with high precision of 95.8% and high recall of 82.6%. With the quality classification in place, 70% of the QCD segmentations achieved a high mean DSC of 0.859 and near-perfect Pearson correlation of 0.97 in estimating the LV myocardial T1 compared to the manual ground truth. Only 30% of the segmentations required manual reviews or corrections.

With a processing speed of 0.39 second per input image estimated in Chapter 3, the use of the QCD framework could lead to real-time parameter extraction in clinical practice and automation of labuor-intensive tasks in large-scale clinical research and trials. This can enable clinicians and healthcare personnel to spend more time with patients rather than performing tedious segmentation and quality control tasks.

Candidate's Contribution

Conceptualisation, Methodology, Software, Experiments, Analysis, Data curation, Literature reviews, Writing - original draft, Writing - review and editing

Chapter 5

Generalisability of the QCD Framework to CMR T1 Maps with Suboptimal Image Quality: a Visual Assessment

5.1 Introduction

In Chapter 4, the QCD framework, which was trained using internal OCMR datasets, has demonstrated excellent generalisability to the previously unseen external UKBB dataset of 2020 T1 maps. Firstly, the QCD framework achieved excellent DSC performance (median of over 0.85) and the highest value for the minimum DSC (0.55), compared with candidate models, demonstrating robustness to outliers. Secondly, the quality control component demonstrated excellent results with a MAE ≤ 0.042. Thirdly, the segmentation quality classification obtained a high precision of 95.8% and high recall of 82.6%.

However, the QCD segmentation and quality control have only been evaluated quantitatively against the manual ground truth by measuring DSC (degree of overlap between the 2 contours), but without visual assessment by a human operator. As shown in Chapter 4 Figure 4.4C and D, evaluation against manual contours using the DSC alone may not always accurately reflect the quality of the automatic segmentations, as the manual contours are subject to interobserver variability even for very experienced operators. Furthermore, quantitative evaluation metrics such as DSC may not measure all aspects of segmentation quality, such as topology of the segmentation [22]. For example, it is still possible for an automatic segmentation to have a false topology despite having achieved a high DSC. Thus, it is desirable to verify whether the automatic QCD segmentations meet other metrics of quality. This can be carried out by expert image analysts to visually evaluate the segmentation quality. Further, the human visual assessment can be used as a reference to evaluate the performance of the quality control component of the QCD framework.

5.1.1 Related Work

Among automatic segmentation methods proposed for CMR T1 mapping [60, 18, 43], only [43] reported visual assessment of segmentation quality. 1500 automatic segmentations were annotated manually for segmentation correctness to develop and evaluate the quality control component [43].

For CMR cine imaging, a number of publications reported various approaches for visual assessment of automated segmentation quality [25, 22, 26, 33]. Examples include manual annotation of segmentation failures [22, 26], a side-by-side comparison of automatic and manual segmentations [25], and subjective scoring of segmentation quality [33].

Similar to some of the automated segmentation methods for CMR cines [26, 33], there is no ground truth segmentation available for the data considered in this chapter, for quantitative evaluation. Thus, it makes practical sense to perform post-hoc visual inspection and manual scoring of segmentation quality.

5.1.2 Objectives

In this chapter, the automatic QCD segmentation was evaluated qualitatively on T1 mapping images available from the UKBB, not used in previous chapters. Manual scoring on segmentation was carried out by an expert observer (SKP) with over 10 years of experience in T1 map analysis. The quality control component (both the DSC prediction and the binary quality classification) of the framework was compared against the manual scores as part of this qualitative evaluation.

5.2 Material and Methods

5.2.1 The UKBB T1 Mapping Dataset

In this chapter, 2266 ShMOLLI native T1 mapping images, acquired for the UKBB, were used as testing data. In contrast to the 2020 T1 maps used in Chapter 4, these data were considered to have suboptimal image quality, with image artefacts affecting more than 1 segment, previously already annotated by another independent experienced image analyst (EL) (see Table 5.1). Given the presence of artefacts, the image analyst (EL) only inserted rough semi-automatically initialised "place-holder" contours to indicate the approximate location of the LV, but did not fine tune to provide gold-standard ground truth. A representative example is shown in Figure 5.1A, in which the contours were not refined to the gold standard, in contrast to the high-quality contours in Chapter 4, as shown in Figure 5.1B. These were different from the high-quality manual contours available in Chapter 4. Thus, in this dataset with suboptimal image quality, it is more appropriate to first deploy the QCD framework to place automatic contours with subsequent human visual assessment of contour quality, rather than to evaluate the automatic contours against these sub-standard manual "place-holder" contours.

5.2.2 The QCD Framework

The same QCD framework, trained in Chapter 3 and evaluated quantitatively for agreement with human contours in Chapter 4, was used for image post-processing in this chapter.

5.2.3 Manual Scoring of Segmentation

The 2266 QCD T1-map segmentations were presented sequentially in a randomised order to an expert image analyst (SKP) via a dedicated software graphical user interface (GUI, Figure 5.2), which I developed in-house using TkInter[1] in Python 3. The GUI displays automatic contours overlaid with the corresponding T1 map in grey

[1]https://wiki.python.org/moin/TkInter

Table 5.1: Factors for rejection of a myocardial segment in T1 map image quality scoring.

Factors for Rejection of a Myocardial Segment in T1 Map Image Quality Scoring
Likely Pathology
Poor planning (Slice level is incorrect, LVOT seen, Apex low, etc.)
Motion (Breathing or Cardiac) artefacts
Mistrigerring
Image Acquisition and Reconstruction (SSFP, phase vortex, fat-water interface, wrap-around, fat partial volume) artefacts
Unidentified artefacts

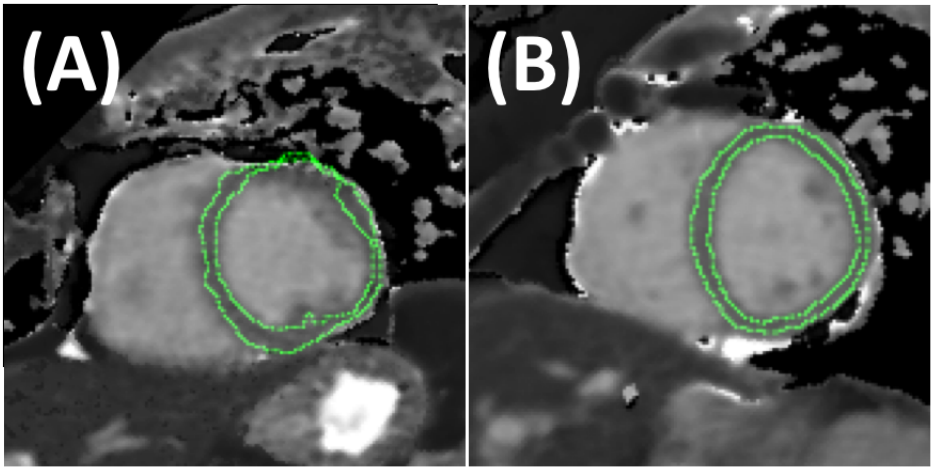


Figure 5.1: Two representative examples of manual contours for (A) Chapter 5, in which the initial semi-automatic contours were not refined to the gold standard, and (B) Chapter 4, in which high-quality manual contours were available.

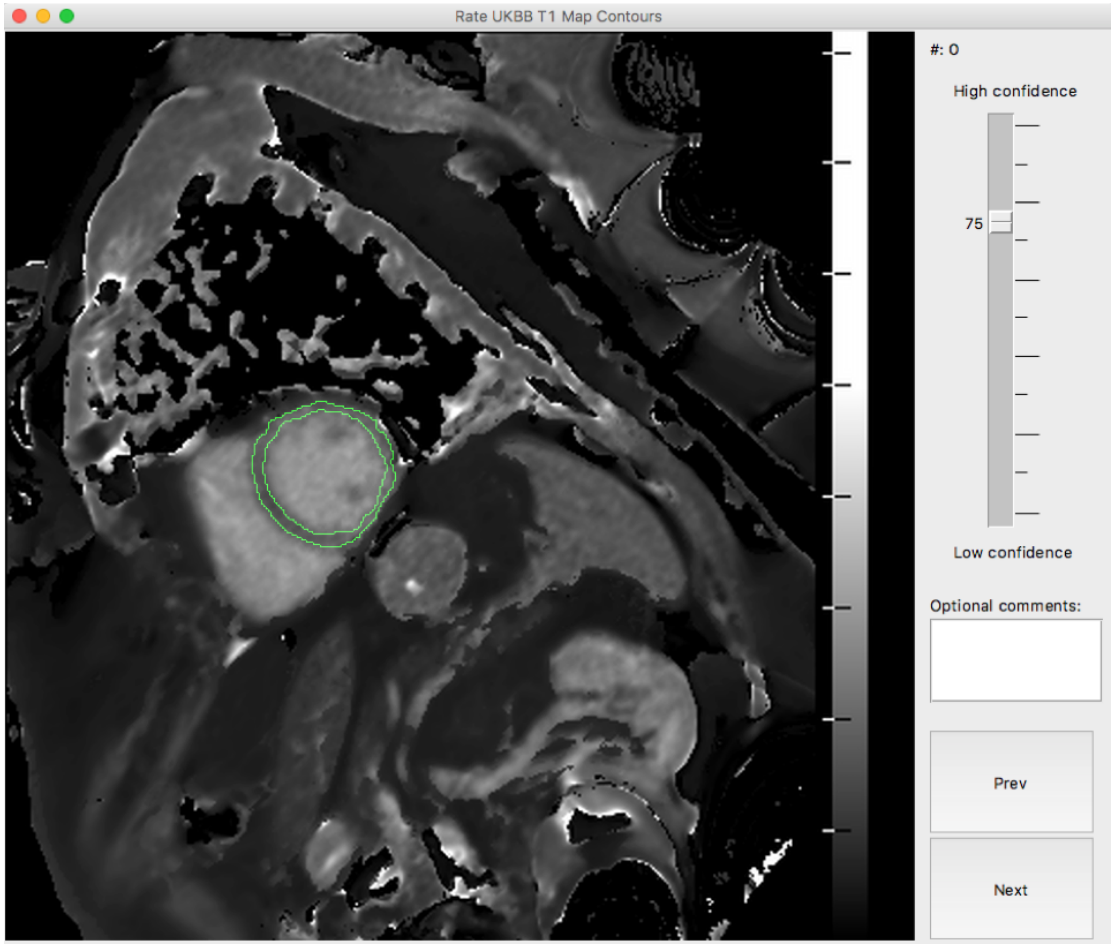


Figure 5.2: Graphical user interface (GUI) displaying a QCD segmentation overlaid onto a T1 map (left), with a slider bar for quality scoring (right). The GUI was developed in-house using TkInter in Python 3.

scale, with a sliding bar on the right-hand side for segmentation quality scoring. An optional text box is available for remarks. Data are saved with a press of the "Enter" key after which next image is presented. Buttons are available as additional means of navigation between images.

The operator was required to use the vertical scoring bar to express their confidence in the accuracy of the automatic contours. The scale of confidence ranges from 0 (the lowest confidence) to 100 (the highest confidence), based on the segmentation location, the shape of the segmentation, and the extent of potential manual correction required. A perfect segmentation is expected to delineate the LV myocardium accu-

Table 5.2: Guideline descriptions for different manual scoring ranges.

Score	Guideline description
90 to 100	Perfect segmentation
80 to 89	Near-perfect segmentation; slight retouch is optional
70 to 79	Small retouch is required at one segment
60 to 69	>1 segment but less than half of the myocardium requires retouching
50 to 59	About half of the myocardium requires retouching
40 to 49	About half of the myocardium requires redrawing
30 to 39	Most of the myocardium requires redrawing
20 to 29	Unusable segmentation; only some overlap with the myocardium
10 to 19	Completely wrong segmentation placement or shape
0 to 9	No segmentation at myocardium

rately, without including the papillary muscles, the blood pool, or any extra-cardiac structures. Such segmentation can be used reliably for clinical studies without any manual retouching. The scoring criteria are listed in Table 5.2, with 10 bands: 0 to 9, 10 to 19, 20 to 29, and so on.

Within the material, a random sample consisting of 227 segmentations was presented to the observer twice, to allow evaluation of intra-observer variability. Pearson correlation (r) and intraclass consistency coefficient (ICC; two-way mixed model) were reported to show whether the repeated scores were consistent. Statistics on absolute differences between the repeated scores were also reported. For the purpose of further analysis, the average values between repeated assessments were used, instead of having two scores for the same segmentation.

5.2.4 Segmentation Quality Classification

The DSC prediction in the QCD framework was extended to segmentation quality classification, similar to Chapters 3 and 4. The manual quality scoring was used as the ground truth quality using two thresholds (80 and 60) to classify automatic segmentations which require varying levels of attention from human operators. A manual score of 80 or above served to indicate a good quality segmentation which does not require manual correction, while a score of below 80 but 60 or above indicated that manual retouching is required for the segmentation. A score of below 60

served to exclude segmentations perceived as completely incorrect or requiring major manual corrections. For each manual score threshold, an optimal threshold for the predicted DSC had to be chosen using the maximum Youden index on a receiver operating characteristic (ROC) curve. Similar to Chapters 3 and 4, further evaluation was performed using binary classification metrics, such as accuracy (ACC), precision (PRC), and recall (REC).

5.3 Results

5.3.1 Manual Scoring of Segmentation

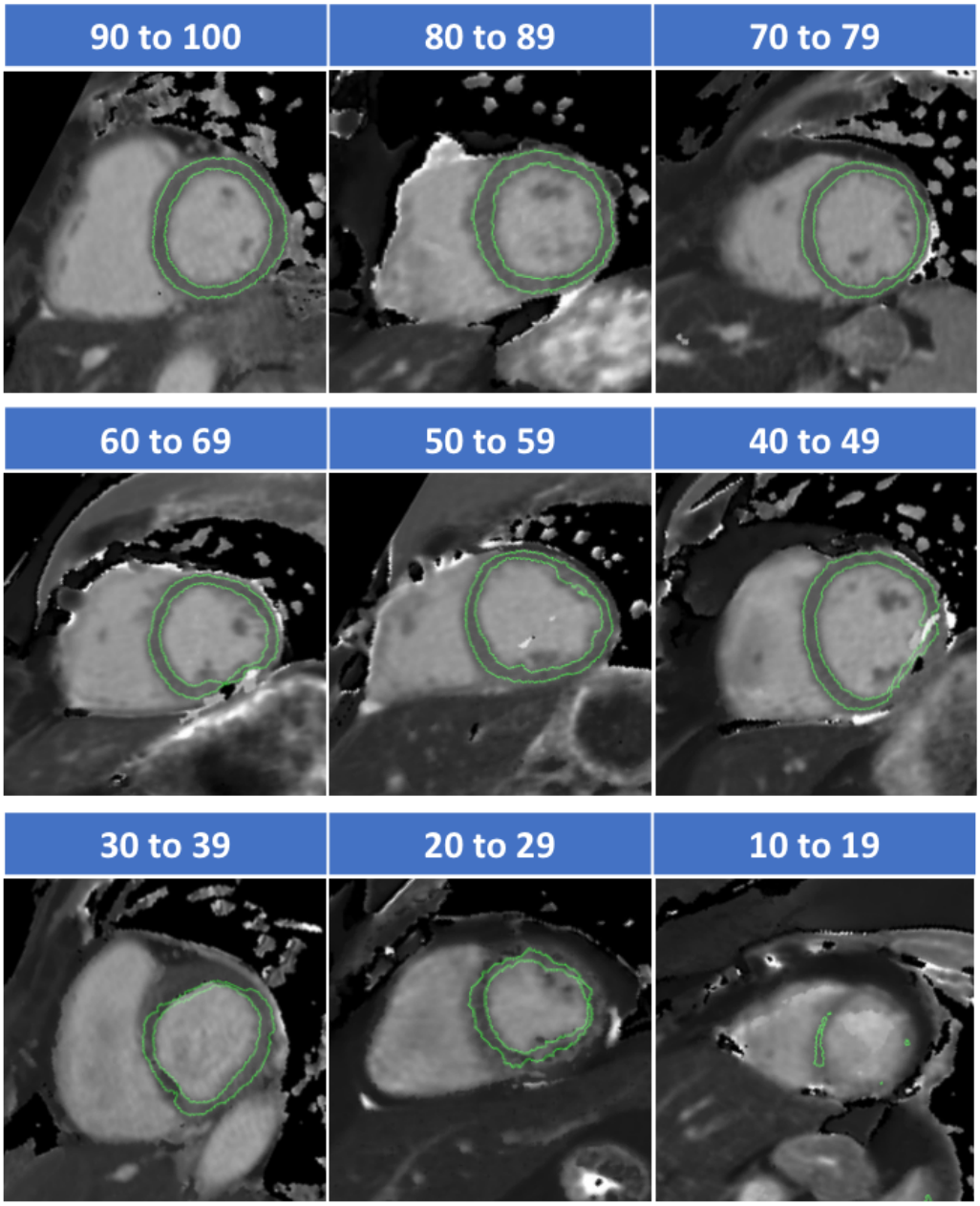


Figure 5.3: Representative examples for different manual score ranges (indicated in blue boxes). As no segmentation has been scored under 10, there is no example for manual score in 0 to 9.

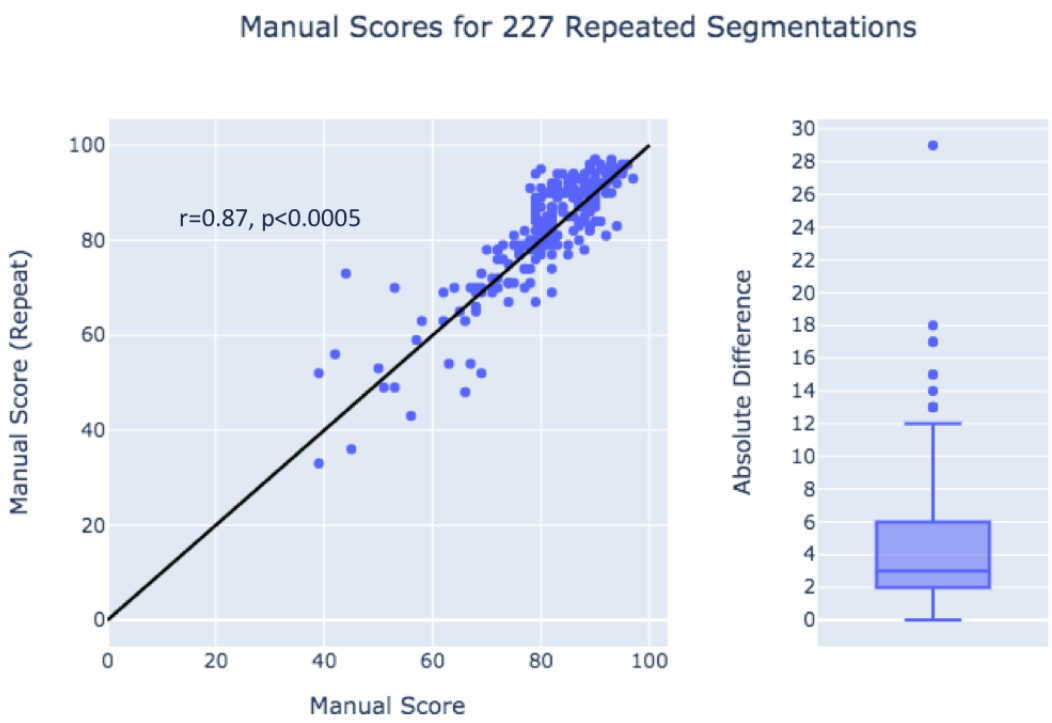


Figure 5.4: Intra-observer variability for manual quality scoring in a selection of 227 repeated T1-map segmentations assessed by a single observer. A scatter plot (left) shows a very good Pearson correlation ($r = 0.87, p < 0.0005$) of the repeated scoring. A box plot (right) shows that the median absolute difference is just 3 on a 100-point scale.

A total of 12 hours and 39 minutes were used to annotate the segmentation quality by a single analyst according to the guidelines described in Table 5.2. Representative examples are shown in Figure 5.3 for different manual scoring ranges: 90 to 100, 80 to 89, 70 to 79 and so on. Since no QCD segmentation was scored under 10, there is no example found for manual scoring range 0 to 9.

The left panel in Figure 5.4 shows a scatter plot (left), in which the pair-wise repeated scores show excellent linear correlation ($r = 0.87, p < 0.0005$). The right panel in Figure 5.4 summarises the distribution of absolute differences between the pair-wise repeated scores, characterised by a small median difference (3). The consistency measured by ICC (two-way mixed model; average measures) was very high at 0.930 ($p < 0.0005$), and the ICC for absolute agreement was also very high at

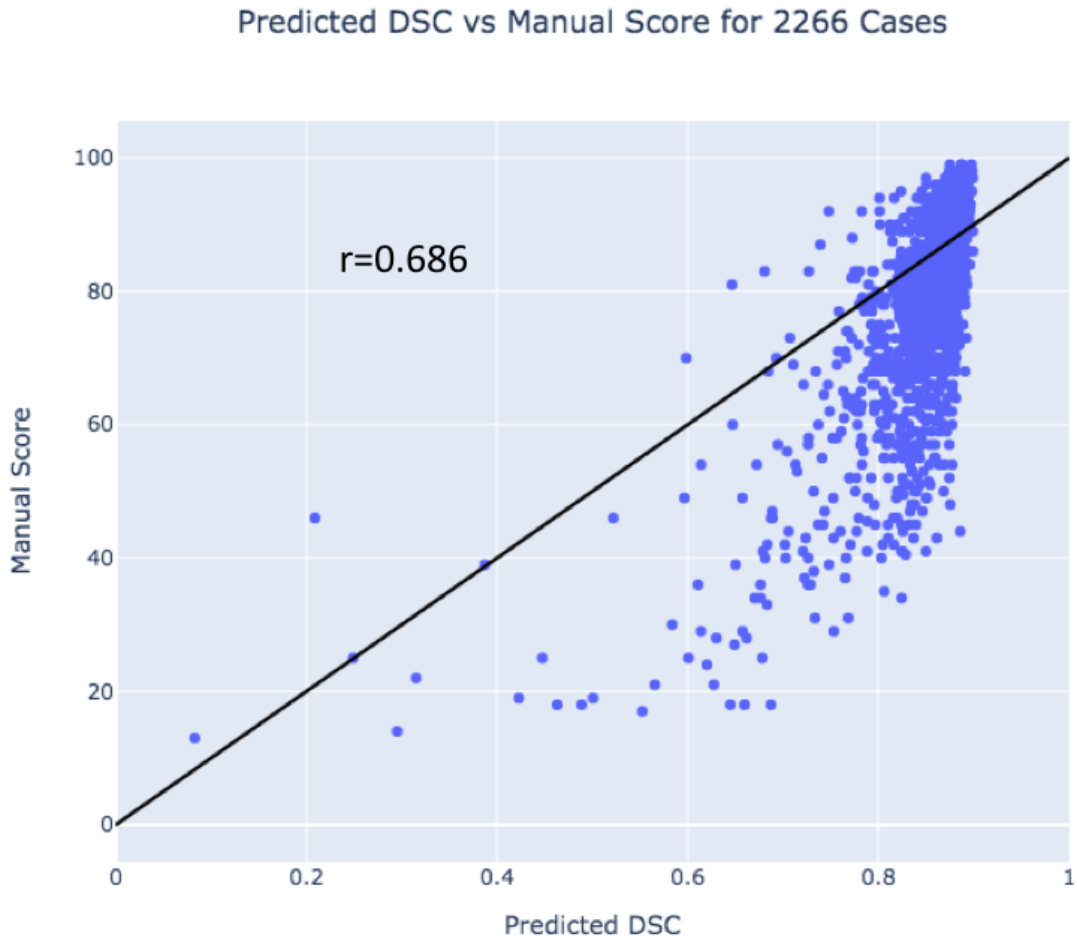


Figure 5.5: Visual assessments of contour quality (x-axis) show clear but non-linear association to the predicted DSCs (y-axis).

0.926 ($p < 0.0005$). These results demonstrate that the manual segmentation quality scoring by the human analyst was consistent for the repeated cases, and thus can be relied upon to evaluate the QCD quality predictors.

The predicted DSC was correlated with the manual scores of segmentation quality as shown in Figure 5.5, with a Pearson correlation of 0.686 ($p < 0.0005$). Hence, the DSC prediction generally agreed with the opinions of the expert image analyst in quantifying segmentation quality, even though the predicted DSCs and the manual scores reflected different aspects of segmentation quality.

The DSC prediction was evaluated for its predictive value to detect selected thresh-

olds of manual scoring for segmentation quality classification. Two manual score thresholds of 80 and 60 were considered. The ROC curve in Figure 5.6A shows that the DSC prediction could detect good quality segmentations (manual score $\geq$ 80), with a high AUC of 0.810 ($p < 0.0005$). The Youden index (red dashed line in Figure 5.6A) on the ROC curve indicated an optimal threshold at 0.860 (green dot in Figure 5.6A), which was close to the previous threshold at 0.866 (purple dot in Figure 5.6A), established in Chapter 4. These two thresholds came with similar specificities (0.673 for the threshold at 0.860; 0.758 for the threshold at 0.866) and sensitivities (0.789 for the threshold at 0.860; 0.690 for the threshold at 0.866), with differences less than 0.1. This may imply that the thresholding scheme is robust despite image quality variation.

The ROC curve in Figure 5.6B shows the detection of cases with manual scores $\geq$ 60, achieving a higher AUC (0.911, $p < 0.0005$), compared to that in Figure 5.6A. With an optimal threshold at 0.844 (green dot in Figure 5.6B), higher specificity (0.851) and sensitivity (0.835) were also achieved. This implies that the DSC prediction was better at detecting cases with manual scores $\geq$ 60 than cases with manual scores $\geq$ 80, and demonstrates that the automatic DSC prediction can be used reliably to filter out potentially problematic cases with lowly-perceived segmentation quality.

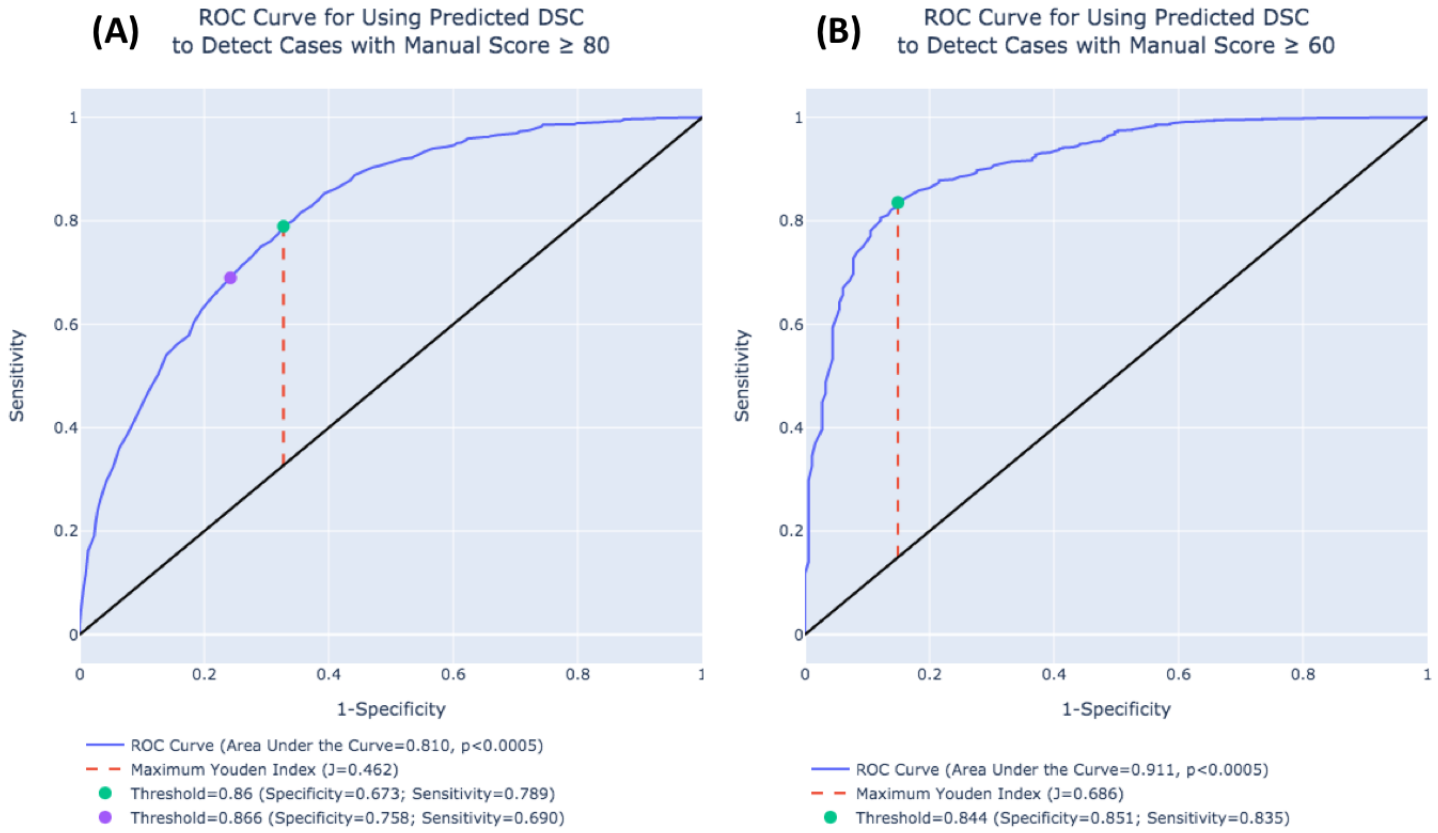


Figure 5.6: Predictive properties for selected thresholds in segmentation quality assessment using predicted DSC. (A) Detection of cases with manual scores ≥ 80, with AUC=0.810 ($p < 0.0005$). An optimal threshold for the predicted DSC was found at 0.860 (green dot) using the maximum Youden index. The optimal threshold for the predicted DSC (0.866; purple dot) found in Chapter 4 is also shown for comparison. (B) Detection of cases with manual scores ≥ 60 with AUC=0.911 ($p < 0.0005$). An optimal threshold for the predicted DSC is found at 0.844 (green dot) using the maximum Youden index as a criterion.

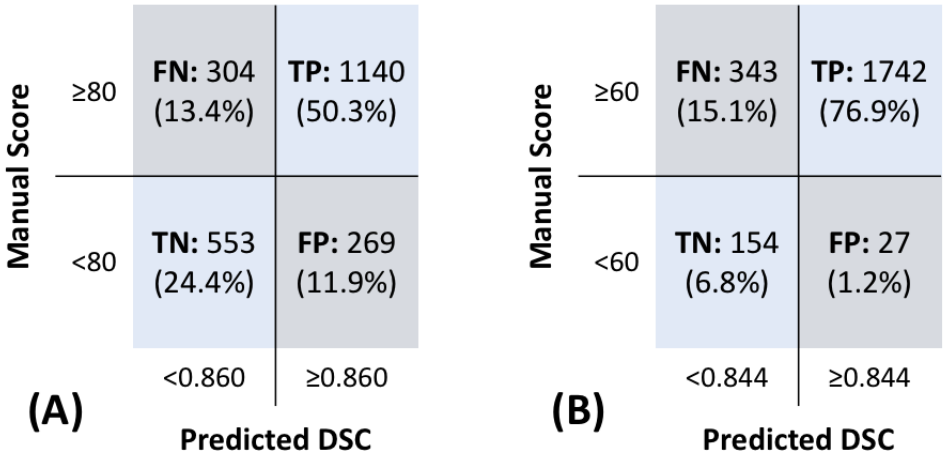


Figure 5.7: Two confusion matrices showing the number of cases for true positive (TP), false positive (FP), false negative (FN), and true negative (TN), respectively, corresponding to the thresholds shown in the ROC curves in Figure 5.6.

With the predicted DSC threshold of 0.860, 50.3% of all cases were detected as true positive, having manual scores ≥ 80 (Figure 5.7A). These QCD segmentations were considered near-perfect quality by the expert image analyst and could be used for further processing reliably without any manual correction. With 24.4% true negative cases, the QCD framework correctly classified the majority of the cases with an accuracy of 74.7%, indicating generally good performance to differentiate segmentation quality at the manual score threshold of 0.860. While 11.9% false positive cases (Figure 5.7A) were found, it is expected that these cases, with high predicted DSC, would not significantly affect the myocardial T1 estimation, as per the results in Chapter 4. Manual ground truth contours would be required to confirm the actual impact on T1 estimation. Furthermore, the classification precision of 80.9% indicates that the vast majority of the predicted positive cases were considered good quality by the expert image analyst, thus the majority of the predicted positive cases were considered reliable for estimating T1 values.

With the predicted DSC threshold of 0.844, the classification achieved a high accuracy of 83.7% (Figure 5.7B), with 76.9% true positive and 6.8% true negative, implying good classification performance. The 76.9% true positive cases would only

require minimal manual retouching at the most, for example to reduce partial volume effects, not requiring time-consuming manual redrawing. The classification precision was very high at 98.5%, with very few false positive cases (only 1.2%). This implies the quality control was successful to screen out the vast majority of the cases with low manual scores below 60 from the positive cases. In addition, the low number of false positive cases also implies that most of the false positive cases in Figure 5.7A were still scored at least 60 by the expert image analyst, potentially with moderate impact on myocardial T1 estimation. Therefore, the quality classification was successful for both predicted DSC thresholds at 0.860 and 0.844, with good overall performance in grading the segmentation quality against human assessment.

Examples of different classification results are shown in Figure 5.8 by category, in relation to both classification systems, with the manual score thresholds at 80 and 60, and the respective predicted DSC thresholds at 0.860 and 0.844. The true positive case (Figure 5.8A) shows a segmentation with a high predicted DSC (0.899), which was consistent with the high manual score (97). The true negative case (Figure 5.8B) shows a segmentation which failed to delineate most of the myocardium. The segmentation was not in an annulus-like shape, and thus scored very low (13) by the expert image analyst. The extremely low DSC prediction (0.083) reflects very well the low manual score.

The false positive example (Figure 5.8C) shows a segmentation which was affected by an image artefact (low intensity pixels). The segmentation included a few extra disconnected pixels in the image artefact in the blood pool (red arrow in Figure 5.8C). These pixels were affecting the topology of the segmentation, and also potentially affecting the myocardium T1 estimation due to inclusion of other structures. The expert image analyst considered this as a critical error, thus gave the segmentation a low score of 52. Interestingly, the predicted DSC was high (0.875), correctly reflecting the very small area actually affected (< 10 pixels relative to the whole segmentation mask ~ 300 pixels). This demonstrates how the DSC prediction and the expert image analyst could reflect differently on the same observation, and highlights

a potential limitation for the DSC prediction, which may not capture other aspects of segmentation quality, such as topology.

The false negative example (Figure 5.8D) highlights the reversed phenomenon, in which a segmentation was considered good quality by the expert image analyst, while a low DSC was predicted by the QCD framework. This example shows a segmentation of a thin myocardium, which was considered good quality by the expert image analyst with a high manual score of 83, as it appeared correctly contoured. Interestingly, the segmentation was predicted a low DSC (0.681) by the QCD framework, due to high disagreement among candidate segmentations, likely because of high expected uncertainty in segmenting the thin myocardium over the partial volume. This demonstrates the possibility that the quality control would flag up cases which potentially face serious errors, while human analysts would only evaluate the correctness of the segmentations, and not reflect potential causes of uncertainty in scoring.

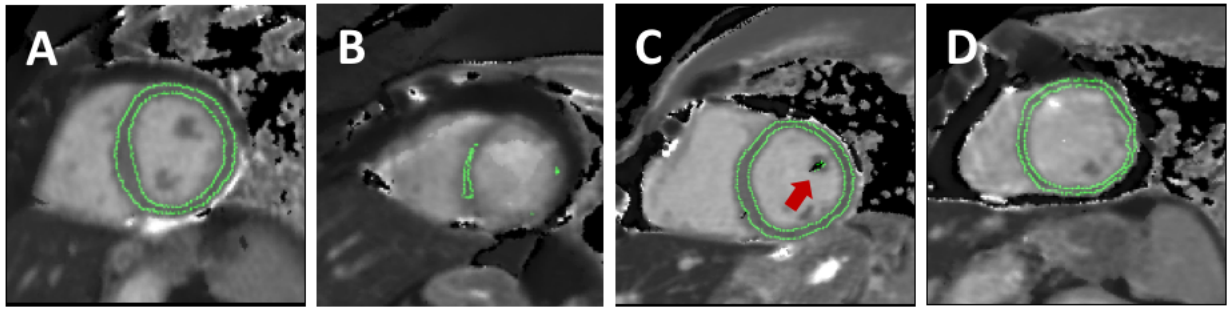
	True Positive	True Negative	False Positive	False Negative
Predicted DSC	0.899	0.083	0.875	0.681
Manual Score	97	13	52	83
QCD Contours				

Figure 5.8: Examples of true positive (A), true negative (B), false positive (C), and false negative (D), showing the QCD segmentations overlaid onto the T1 maps with corresponding predicted DSCs and manual scores. The red arrow in (C) indicates spurious pixels of the segmentation in the image artefact.

5.4 Discussion

The QCD framework was successfully applied to an unseen external dataset of 2266 T1 maps with suboptimal image quality available from the UKBB, which do not have ground truth manual contours. The automatic QCD segmentations were subsequently visually assessed by a reliable expert observer, who had demonstrated minimal intra-observer variability. Comparison with the human visual assessment showed that the QCD framework is able to select high quality segmentations with high accuracy and precision, potentially saving manual effort to inspect or correct segmentations.

5.4.1 Segmentation Quality Classification

The QCD framework has demonstrated promising capability to classify the segmentation quality correctly in relation to the manual threshold at 80. The framework achieved a high accuracy of over 75%. A vast majority of the predicted positive cases were expected to have reliable segmentations for myocardial T1 estimation, as indicated by the classification precision of over 80%. Besides the good classification performance, a clear advantage of the QCD framework over manual visual assessment of segmentation quality is the high speed performance estimated at 0.39 second per image for both segmentation and quality control, over 45 times faster than the expert image analyst (about 18 seconds per image only for visual assessment).

5.4.2 Visual Assessment

The visual assessment provided useful insights into the QCD framework, especially the DSC prediction component. A non-linear association of the visual assessment scoring and the DSC prediction was revealed, demonstrating agreement and some discrepancies between opinions of the expert human observer and the DSC prediction.

This qualitative evaluation complemented the quantitative approach in Chapter 4. Visual assessment was not performed in Chapter 4, as the manual ground truth segmentations were available for quantitative comparison with the QCD segmentations. Furthermore, the QCD framework successfully predicted good quality segmentations

with excellent agreement to the ground truth in estimating myocardial T1. Thus, there was little additional benefit to perform additional visual assessment in Chapter 4.

5.4.3 Comparison with Related Work

As discussed in Chapter 4, one advantage of the QCD framework is cheaper labour and time cost for training, compared to the recently-proposed segmentation quality control for CMR T1 mapping [43]. In this chapter, visual quality scoring for the QCD segmentation had the advantage of capturing the diverse variety of segmentation quality on a scale of 0 to 100, according to the guideline described in Table 5.2. This allowed more flexibility and precision in evaluation, especially on the boundary cases. In contrast, the most recent method took a different approach by giving binary annotations of segmentation correctness, for which the criteria were not stated [43]. This approach may not share the same advantage. Moreover, the work presented in this chapter had a clear guideline in place to allow more consistent manual annotations. This is desirable especially when more annotations are required in the future. Therefore, the approach for the visual quality assessment taken in this work is more advantageous than the most recent method for CMR T1 mapping.

5.4.4 Limitations and Future Work

A direct comparison on quality classification performance between the QCD framework and the most recent method [43] may not be meaningful, as the two were visually evaluated in different approaches as described in Section 5.4.3, and using different parts of the UKBB data for the developments and the evaluations. It will be beneficial in the future to standardise evaluation criteria and datasets, allowing head-to-head comparison among state-of-the-art methods.

In this chapter, the visual segmentation quality assessment was carried out by a single observer. Although the evaluation of intra-observer variability showed that the scoring was consistent, the results could still be specific to the single observer. It is possible that another expert image analyst may hold a different opinion regarding

the quality scoring, and inter-observer variability remains to be assessed in future. By building upon the groundwork laid in this chapter, large-scale validation of QCD segmentations can be carried out with multiple observers in the future on this material and subsets of the target material available from the UKBB.

For the DSC prediction, a limitation has been revealed. As shown in Figure 5.7C, the predicted DSC was high (0.875) even though the segmentation included the image artefact in the blood pool. The inclusion of the image artefact only accounted for 2-3%, in terms of pixel count, of the whole segmentation mask. Thus, the DSC prediction can be insensitive to such small inclusion if the error does not affect the overall agreement among candidate segmentations. To overcome this limitation, the QCD framework assessments need further work to include distance-based or topology-based segmentation quality metrics, and MR image quality assessments. Topological analysis can detect and remove segmentations with spurious pixels and false topologies. In addition, the results of training another network (or the QCD network ensembles) to predict human scores as an additional fail-safe could be employed. Given that this work is dedicated to DSC performance prediction using the QCD, rather than the final product to process the UKBB data, these additional heuristic solutions were not considered here, but is in planned future work.

5.4.5 Conclusion

The QCD framework have been applied to a challenging task of segmenting over 2000 T1 maps which have not been manually contoured for ground truth, due to suboptimal image quality as adjudicated by an experienced human analyst. Subsequently, the automatic QCD segmentations were visually evaluated by a second image analyst for segmentation quality. Despite the suboptimal image quality, the results have shown that over 60% of the segmentations were considered near-perfect quality, and less than 10% would require manual redrawing, demonstrating the robustness of the QCD segmentation.

The visual assessment scores have been compared with the DSC prediction and segmentation quality classification of the QCD framework. The results showed that

the DSC predictions can rapidly (in less than half a second) reflect the opinion of the human image analyst, with a non-linear association. The segmentation quality classification demonstrated high accuracy and precision for identifying cases which would not require manual redrawing, and excellent ability to filter out the cases with serious problems. While demonstrating good performance, the QCD framework is flexible to allow further improvement by incorporating distance-based or topology-based segmentation quality metrics for better quality control accuracy and efficiency. This could pave the way for efficient processing of a larger-scale datasets of 100,000 T1 maps available from the UKBB in the future, with a significant proportion of the manual work saved without sacrificing T1 accuracy.

Candidate's Contribution

Conceptualisation, Methodology, Software, Experiments, Analysis, Data curation, Literature reviews, Writing - original draft, Writing - review and editing

Chapter 6

Summary and Future Work

6.1 Summary

With increasing use of large-scale CMR imaging datasets for investigating cardiovascular diseases, it is desirable to develop reliable automatic algorithms for efficient imaging data analysis and interpretation. Deep learning has been a popular method of choice for automating CMR image analysis, especially for image segmentation. However, even the current state-of-the-art segmentation algorithms can still fail, potentially leading to misestimation of clinical parameters against the best interest of medical practitioners and their patients. There is a rapidly increasing interest in provision of accurate automatic quality control mechanisms with the image segmentation for clinical use.

Deep ensemble methods have been proposed to exploit agreement among multiple deep neural networks to estimate prediction uncertainty, mainly for image classification tasks. It has been shown that deep ensembles can outperform other popular uncertainty estimation methods, such as Monte Carlo-based Bayesian neural networks, in estimating prediction uncertainty, especially in terms of generalisation. However, there is limited exploration of using deep ensembles to segmentation quality metrics for image segmentation in CMR imaging. Furthermore, the state-of-the-art methods have not progressed to suggest that inherent DSC predictions can be used to improve segmentation robustness.

In Chapter 2, I introduced the novel idea of the proposed quality control-driven (QCD) approach for automatic aortic lumen segmentation of $\sim$ 5000 CMR short-axis cine image sequences, available from the UKBB. The QCD framework successfully performed both image segmentation and quality control tasks. Building upon the deep ensemble method, the QCD framework made use of multiple linear regression to predict image-wise segmentation quality metrics (DSC in this work) for each of the members of the ensemble. Based on such predictions, the QCD framework was able to compare multiple candidate segmentations and select the final, most optimal output on-the-fly. The QCD framework achieved the best segmentation performance

compared to the individual candidate models, with near-perfect agreement with the manually-validated ground truth in estimating aortic lumen area.

In Chapter 3, I successfully demonstrated that the QCD framework can be easily adapted to contour and quality control the segmentation of a different CMR imaging technique – native T1 mapping (ShMOLLI method), and of a more complex anatomy – the LV myocardium in short-axis. The QCD framework was trained on manually contoured T1 maps, available internally in our unit (OCMR). After training, the QCD framework achieved top performance for LV segmentation, and near-perfect agreement with the manual ground truth in estimating myocardial T1, with a Pearson correlation of 0.99 and a mean absolute percentage error of about 1%. Furthermore, I extended the DSC prediction of the QCD framework by implementing a simple thresholding scheme to perform binary classification of segmentation quality, with an excellent accuracy of 99%. The framework achieved a rapid processing time of less than half a second per T1 map, making it suitable for real-time clinical applications. This chapter highlighted the desirable properties of the QCD framework: adaptability across imaging techniques and anatomical structures, highly accurate image segmentation and quality control, near-perfect clinical parameter (T1 value) estimation compared with the manual ground truth, and excellent speed suitable for practical processing of the target medical images.

In Chapter 4, I tested the generalisability of the QCD framework in a large volume of unseen external dataset of T1 maps (ShMOLLI method). These data accounted for about half of the $\sim$ 5000 pilot datasets available from the UKBB. The T1 maps used had been manually selected for good to excellent image quality, prior to presentation to the QCD framework for the automatic segmentation task. Due a different version of CMR acquisition protocol used for the UKBB T1 maps, there were changes in the segmentation performance of candidate models, relative to Chapter 3. Despite that, the QCD framework was the only model remaining in the top 3 ranking among candidate models for best segmentation performance in both Chapters 3 and 4, demonstrating its excellent robustness to different datasets. Additionally, I extended the QCD framework to include segmentation quality classification with a thresholding

scheme, and the classification demonstrated excellent capacity to detect good quality segmentations, which estimated myocardial T1 values very close to the ground truth. On the other hand, segmentations with predicted DSC below the threshold had lower agreement in estimating myocardial T1 values. This further demonstrated the success of the segmentation quality classification in identifying automatic segmentations that are acceptable for reliable T1 estimation. Thus, the QCD framework demonstrated excellent generalisability and robustness in automatic segmentation, quality control, and T1 estimation, even when applied to a large-scale external dataset acquired with a different set of scanning parameters. Considering the ever-advancing CMR technology, this generalisability of the QCD framework is both robust and highly desirable, especially for implementation in real-world clinical applications.

In Chapter 5, I tested the robustness of the QCD framework in segmenting a large-scale dataset with suboptimal image quality. T1 maps with suboptimal image quality were used. In view of the lack of ground truth segmentation, the quality of the automatic QCD contours were evaluated by visual scoring by an expert observer. Over half of the QCD automatic contours were considered to have near-perfect contour quality on visual assessment, demonstrating the robustness of the QCD framework in segmenting even datasets with suboptimal image quality. The quality control component of the QCD framework also demonstrated a strong association with the human scores, and very good performance in classifying segmentation quality. Thus, the QCD framework is highly promising for real-world clinical applications and large-scale imaging studies, including both optimal and suboptimal image data.

6.2 Future Directions

The QCD framework has been developed and tested only on T1 maps acquired using shortened modified Look-Locker inversion recovery (ShMOLLI) rather than modified Look-Locker inversion recovery (MOLLI), which may be more widely used. MOLLI was proposed in 2004 for measuring T1 values across the myocardium by acquiring 11 T1-weighted samples [83]. However, a long breath hold of 17 heart beats is required for a scan using MOLLI, presenting challenges to elderly population and patients with compromised cardiac and respiratory functions [61]. By acquiring only 7 T1-weighted samples, ShMOLLI was proposed in 2010 to reduce the breath hold time to only 9 heart beats [61]. Despite having a lower signal-to-noise ratio due to fewer T1-weighted samples, the shorter breath hold requirement makes ShMOLLI a more practical alternative to MOLLI for patients who find difficulties in long breath holds [61]. Furthermore, large-scale imaging studies such as the UKBB have selected ShMOLLI as the technique of choice for T1 mapping [3]. This can potentially encourage wider adoption of ShMOLLI for research and clinical applications, and offer exciting opportunities for the QCD framework to process large databases of unannotated ShMOLLI data. Thus, ShMOLLI T1 mapping has been chosen as the focus for the QCD framework in this thesis. Subject to the availability of annotated data, the QCD framework can be trained solely on MOLLI T1 mapping as an extension in the future for wider clinical usage. In addition, future work can investigate possibility of exploiting raw T1-weighted samples from ShMOLLI or MOLLI for further improvement in segmentation, as raw samples may contain more information.

In clinical practice, under-segmentation of the myocardium is preferred by clinicians to avoid partial volume. Inclusion of partial volume into the myocardial segmentation mask has been identified as the primary source of potential error in estimating global T1 values [65]. It has been found that inflated myocardial contours which include pixels in the neighboring tissues, especially the blood pool, can overestimate the myocardial T1 values by up to 6.9%, and increase variability [65]. It is recommended that image analysts can erode the myocardial borders by 1 pixel to obtain a mid-wall

myocardial segmentation mask for consistent inter and intraobserver agreement when the myocardial coverage is not considered critical [65]. Therefore, further investigation on eroding automatic segmentation may lead to potential improvement in accuracy of T1 estimation. Future implementation may include a graphical user interface to allow flexible adjustment of the myocardial segmentation thickness for different clinical needs, and enable batch-processing for precise and customisable processing of T1 maps, suitable for efficient clinical workflows.

In this thesis, the QCD framework focused on predicting DSC for quality control. However, DSC is known to have some limitations. It has been shown that DSC generally favours larger-sized structures [84, 85]. Thus, it may not be fair to use DSC as a quality metric to directly compare segmentations for the mid-ventricle with smaller segmentations for the apex. For future work, it may be beneficial to set different DSC thresholds independently for basal, mid-ventricular, and apical slices to better differentiate segmentation quality.

Another limitation is that DSC only measures the overlap with the ground truth segmentation, but may not capture other aspects of segmentation quality, such as the shape and the topology of the contours. It is possible for a topologically or anatomically incorrect segmentation, e.g. having disconnected components, to obtain a high DSC. As the predicted DSC has a linear relationship with the observed DSC (shown in Figure 3.6 in Chapter 3), it can suffer the same limitation. Figure 5.8 in Chapter 5 shows an example in which there were spurious disconnected segmentation components outside the myocardium despite obtaining a high predicted DSC, resulting in a false positive for the segmentation quality classification. To mitigate this problem in the short term, connected component analysis can be introduced to detect disconnected pixels outside the myocardium. This can be implemented alongside the DSC prediction to quality control each candidate segmentation. Erroneous candidates, with incorrect shapes or topologies, can be automatically flagged for inspection. This can reduce the negative impact brought about by the small proportion of observed erroneous segmentations, and improve the reliability of clinical parameter estimation (e.g. T1 values).

In the future, it may be desirable to have integral mechanisms to detect other aspects of segmentation quality. Other quality metrics, such as Hausdorff distance, can also be implemented to capture distance-based quality information to further improve the quality control mechanism. In addition to predicting quantitative segmentation metrics, more advanced non-linear regression models, such as CNN, can be trained to predict visual assessment annotations, similar to [43]. However, this can come with a hefty cost of manual labour and time to prepare a large-scale annotated dataset via visual assessment, similar to that of Chapter 5. Future research is required to select effective quality measures to incorporate into the QCD framework, further refining it for clinical grade applications.

Incorporation of the goodness-of-fit map (also known as the R^2 map) into the QCD framework can potentially enable image quality control as part of the T1 map processing pipeline. The R^2 map is available along a ShMOLLI T1 map to show pixel-by-pixel how well each T1 value fits the recovery curve model [63]. As R^2 is reduced in the presence of off-resonance frequency shifts, motions, or other error sources, inspection of the R^2 map can serve as a quality control process for T1 map image quality [63]. Myocardial segments with presence of low R^2 values are usually excluded from T1 estimation to avoid inaccuracy [63]. For future work, automatic detection of low R^2 values can offer a one-stop solution for T1 map processing integrating both segmentation and image quality control.

Beyond predicting segmentation quality, it is desirable to predict the quality of the clinical parameter estimation, such as the LV myocardial T1 for CMR T1 mapping. Similar to prediction of segmentation quality, this can be implemented by comparing the estimated T1 values among multiple candidates to predict the expected degree of error. As clinical parameter estimation is influenced by the quality of the segmentation, further research can be done to incorporate quality prediction of estimated clinical parameters into the on-the-fly selection of the final most-optimal segmentation in the QCD framework, improving quality of both the segmentation and clinical parameter estimation.

Another future research direction may look into improving the QCD framework by selecting a diverse set of segmentation methods and techniques to combine candidate segmentations. It has been shown that diversity among candidate segmentations could be important for robust quality control; further research can be done to assess potential benefits of incorporating different segmentation methods, for example anatomically-constrained neural network, which utilises an additional auto-encoder to learn to impose anatomical constraints [86]. The flexibility to incorporate any prior and future segmentation models is one of the major advantages of the QCD framework, to further improve both accuracy and reliability of the segmentation and the quality control, that can adapt to ever-evolving medical image analysis research.

The software implementation of the QCD framework in image processing pipelines of clinical MR scanners has been on going and will continue in the future, in collaboration with industrial partners. As the QCD framework can achieve a rapid processing speed of 0.39 second per image (reported in Chapter 3), it will benefit clinical imaging analysis with reliable real-time segmentation, clinical parameter extraction, and quality control. Future work is pending to further validate the capability of the framework to handle variabilities among different MRI scanners and acquisition parameters.

Further testing of the QCD framework on large-scale pathological CMR data is pending. For the T1 mapping application, the QCD framework was developed with clinical CMR imaging data, acquired from both cardiac patients and healthy controls in OCMR. Subsequently, the QCD framework was tested on a small unseen subset of the OCMR material (Chapter 3), followed by large-scale testing (Chapter 4) and quality inspection (Chapter 5) using the UKBB data, acquired from mostly healthy population [3]. While the QCD framework has demonstrated overall excellent performance on the UKBB data, successful translation to large-scale pathological data is still pending. Pathologies can affect myocardial T1 and even cardiac morphology, presenting challenges for automatic segmentation algorithms, which could fail to generalise when not adequately trained [11]. Subsequently, this may also present a challenge for quality control as the number of poor-quality segmentations may significantly increase. Thus, evaluation of the QCD framework on large-scale pathological

datasets is important to ensure reliability for clinical applications. For future evaluation, the Hypertrophic Cardiomyopathy Registry (HCMR) can provide large-scale T1 mapping data acquired from 2750 hypertrophic cardiomyopathy patients across 44 sites [5]. In addition to native T1 maps, HCMR also provides post-contrast T1 maps for measurement of extracellular volume [5]. Segmentation of post-contrast T1 mapping images is challenging as the myocardial borders usually appear ill-defined. Thus, it is vital to couple automated segmentation with robust quality control to detect potential inaccuracies. The HCMR dataset can potentially prove the reliability of the QCD framework for both segmentation and quality control under complex real-world clinical imaging settings.

The most impactful target is to deploy the QCD framework to process the UKBB large-scale imaging dataset of 100,000 subjects to be made available by 2021. The QCD framework can provide automatic segmentation and estimation of clinical parameters, with prediction of expected quality, appropriately flagging cases requiring manual inspection and correction when necessary. The improvements for the QCD framework that I have suggested in this section can be an on-going effort, in parallel with the deployment to the UKBB data, allowing progressive decrease of the vast amount of manual intervention required.

6.3 Conclusion

In conclusion, I have presented, in this thesis, the novel QCD framework (with a patent application filed) for image segmentation of CMR imaging data with inherent quality control. Based on the concept of deep ensembles consisting of U-nets with different depths, the QCD framework can accurately predict segmentation DSC, using additional regression models. I have successfully applied the QCD framework to two different imaging techniques (cine aortic cross-sections, and ShMOLLI native T1 mapping), in two topologically different cardiac anatomical structures (the circular aortic sections, and the annular LV myocardium shapes). Furthermore, I have also shown that a properly trained QCD framework can reliably address large-scale external unseen datasets from the UKBB, and remain robust in segmentation performance, even in datasets with suboptimal image quality and without ground truth contours. The DSC quality prediction can be extended to segmentation quality classification by implementing a thresholding scheme. This has vast potential to save much of human work by automatically selecting good quality segmentations to reduce manual effort. These excellent results demonstrate in principle that the QCD framework can be used for various other CMR imaging applications, with potential to reach applications beyond image processing, as a general classification solution characterised by a novel robust inherent quality control mechanism.

Bibliography

[1] WHO, "WHO — The top 10 causes of death," 2017.

[2] British Heart Foundation, "CVD Statistics - BHF UK Factsheet," 2018.

[3] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, A. A. Young, S. Garratt, T. Peakman, J. Sellors, R. Collins, and S. Neubauer, "Imaging in population science: Cardiovascular magnetic resonance in 100,000 participants of UK Biobank - Rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, p. 46, dec 2013.

[4] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt, R. Collins, S. Piechnik, and S. Neubauer, "UK Biobank's cardiovascular magnetic resonance protocol," *Journal of Cardiovascular Magnetic Resonance*, vol. 18, p. 8, feb 2016.

[5] C. M. Kramer, E. Appelbaum, M. Y. Desai, P. Desvigne-Nickens, J. P. DiMarco, M. G. Friedrich, N. Geller, S. Heckler, C. Y. Ho, M. Jerosch-Herold, E. A. Ivey, J. Keleti, D. Y. Kim, P. Kolm, R. Y. Kwong, M. S. Maron, J. Schulz-Menger, S. Piechnik, H. Watkins, W. S. Weintraub, P. Wu, and S. Neubauer, "Hypertrophic Cardiomyopathy Registry: The rationale and design of an international, observational study of hypertrophic cardiomyopathy," *American Heart Journal*, vol. 170, pp. 223–230, aug 2015.

[6] S. E. Petersen, M. M. Sanghvi, N. Aung, J. A. Cooper, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, S. K. Piechnik, and

S. Neubauer, "The impact of cardiovascular risk factors on cardiac structure and function: Insights from the UK Biobank imaging enhancement study," *PLoS ONE*, vol. 12, p. e0185114, oct 2017.

[7] C. Petitjean and J. N. Dacher, "A review of segmentation methods in short axis cardiac MR images," apr 2011.

[8] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, pp. 155–195, apr 2016.

[9] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation," *Medical Image Analysis*, vol. 30, pp. 120–129, may 2016.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," dec 2017.

[11] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep Learning for Cardiac Image Segmentation: A Review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1, 2012.

[13] D. Shen, G. Wu, and H. I. Suk, "Deep Learning in Medical Image Analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, jun 2017.

[14] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," *Medical Image Analysis*, vol. 30, pp. 108–119, may 2016.

[15] P. V. Tran, "A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI," *arXiv*, apr 2016.

[16] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, may 2015.

[18] A. S. Fahmy, H. El-Rewaidy, M. Nezafat, S. Nakamori, and R. Nezafat, "Automated analysis of cardiovascular magnetic resonance myocardial native T1 mapping images using fully convolutional neural networks," *J Cardiovasc Magn Reson (JCMR)*, vol. in-press, p. 7, dec 2018.

[19] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10663 LNCS, pp. 120–129, Springer Verlag, 2018.

[20] H. Zheng, Y. Zhang, L. Yang, P. Liang, Z. Zhao, C. Wang, and D. Z. Chen, "A New Ensemble Learning Framework for 3D Biomedical Image Segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5909–5916, jul 2019.

[21] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, nov 2020.

[22] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. M. Rohe, X. Pennec, M. Sermesant, F. Isensee, P. Jager, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Isgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. M. Jodoin, "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[23] A. Suinesiaputra, D. A. Bluemke, B. R. Cowan, M. G. Friedrich, C. M. Kramer, R. Kwong, S. Plein, J. Schulz-Menger, J. J. Westenberg, A. A. Young, and E. Nagel, "Quantification of LV function and mass by cardiovascular magnetic resonance: Multi-center variability and consensus contours," *Journal of Cardiovascular Magnetic Resonance*, vol. 17, p. 63, jul 2015.

[24] V. Carapella, H. Puchta, E. Lukaschuk, C. Marini, K. Werys, S. Neubauer, V. M. Ferreira, and S. K. Piechnik, "Standardized image post-processing of cardiovascular magnetic resonance T1-mapping reduces variability and improves accuracy and consistency in myocardial tissue characterization," *International Journal of Cardiology*, vol. 298, pp. 128–134, 2020.

[25] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert, "Recurrent Neural Networks for Aortic Image Sequence Segmentation with Sparse Annotations," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11073 LNCS, pp. 586–594, 2018.

[26] L. Biasiolli, E. Hann, E. Lukaschuk, V. Carapella, J. M. Paiva, N. Aung, J. J. Rayner, K. Werys, K. Fung, H. Puchta, M. M. Sanghvi, N. O. Moon, R. J. Thomson, K. E. Thomas, M. D. Robson, V. Grau, S. E. Petersen, S. Neubauer, and S. K. Piechnik, "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, feb 2019.

[27] Y. Xu, P. Kavanagh, M. Fish, J. Gerlach, A. Ramesh, M. Lemley, S. Hayes, D. S. Berman, G. Germano, and P. J. Slomka, "Automated quality control for segmentation of myocardial perfusion SPECT," *Journal of Nuclear Medicine*, vol. 50, pp. 1418–1426, sep 2009.

[28] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady, "Evaluating segmentation error without ground truth.," in *Medical Image Computing and Computer Assisted Intervention*, vol. 15, pp. 528–36, 2012.

[29] B. Audelan and H. Delingette, "Unsupervised Quality Control of Image Segmentation Based on Bayesian Learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11765 LNCS, pp. 21–29, Springer, oct 2019.

[30] S. Wang, G. Tarroni, C. Qin, Y. Mo, C. Dai, C. Chen, B. Glocker, Y. Guo, D. Rueckert, and W. Bai, "Deep Generative Model-based Quality Control for Cardiac MRI Segmentation," jun 2020.

[31] 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*, vol. 36, pp. 1597–1606, feb 2017.

[32] R. Robinson, V. V. Valindria, W. Bai, H. Suzuki, P. M. Matthews, C. Page, D. Rueckert, and B. Glocker, "Automatic quality control of cardiac MRI segmentation in large-scale population imaging," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10433 LNCS, pp. 720–727, Springer Verlag, sep 2017.

[33] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee,

V. Carapella, Y. J. Kim, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, P. M. Matthews, D. Rueckert, and B. Glocker, "Automated Quality Control in Image Segmentation: Application to the UK Biobank Cardiac MR Imaging Study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21, p. 18, dec 2019.

[34] R. Robinson, O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, B. Kainz, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, D. Rueckert, and B. Glocker, "Real-Time Prediction of Segmentation Quality," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11073 LNCS, pp. 578–585, Springer Verlag, sep 2018.

[35] R. Robinson, O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, B. Kainz, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, D. Rueckert, and B. Glocker, "Subject-level Prediction of Segmentation Failure using Real-Time Convolutional Neural Nets," in *Medical Imaging with Deep Learning*, no. Midl 2018, pp. 3–5, 2018.

[36] T. DeVries and G. W. Taylor, "Leveraging Uncertainty Estimates for Predicting Segmentation Quality," 2018.

[37] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Inherent brain segmentation quality control from fully convnet monte carlo sampling," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11070 LNCS, pp. 664–672, apr 2018.

[38] A. G. Roy, S. Conjeti, N. Navab, and C. Wachinger, "Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, nov 2019.

[39] H. P. Do, Y. Guo, A. J. Yoon, and K. S. Nayak, "Accuracy, uncertainty, and adaptability of automatic myocardial ASL segmentation using deep CNN," *Magnetic Resonance in Medicine*, vol. 83, pp. 1863–1874, may 2020.

[40] M. Ng and G. A. Wright, "Estimating Uncertainty in Neural Networks for Segmentation Quality Control," in *32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada*, no. Nips, pp. 3–6, 2018.

[41] S. A. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. Ali Eslami, D. J. Rezende, and O. Ronneberger, "A probabilistic U-net for segmentation of ambiguous images," in *Advances in Neural Information Processing Systems*, vol. 2018-Decem, pp. 6965–6975, jun 2018.

[42] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu, "PHiSeg: Capturing Uncertainty in Medical Image Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11765 LNCS, pp. 119–127, jun 2019.

[43] E. Puyol-Antón, B. Ruijsink, C. F. Baumgartner, P.-G. Masci, M. Sinclair, E. Konukoglu, R. Razavi, and A. P. King, "Automated quantification of myocardial tissue characteristics from native T1 mapping using neural networks with uncertainty-based quality-control," *Journal of Cardiovascular Magnetic Resonance*, vol. 22, p. 60, jan 2020.

[44] R. Jena and S. P. Awate, "A Bayesian Neural Net to Segment Images with Uncertainty Estimates and Good Calibration," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11492 LNCS, pp. 3–15, Springer, Cham, jun 2019.

[45] B. Lakshminarayanan, A. Pritzel, and C. B. Deepmind, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Neural Information Processing Systems (NIPS)*, pp. 6405–6416, 2017.

[46] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep Ensembles: A Loss Landscape Perspective," 2020.

[47] S. Lee, S. Purushwalkam, M. Cogswell, V. Ranjan, D. Crandall, and D. Batra, "Stochastic multiple choice learning for training diverse deep ensembles," in *Advances in Neural Information Processing Systems*, pp. 2127–2135, 2016.

[48] E. Hann, L. Biasiolli, Q. Zhang, I. A. Popescu, K. Werys, E. Lukaschuk, V. Carapella, J. M. Paiva, N. Aung, J. J. Rayner, K. Fung, H. Puchta, M. M. Sanghvi, N. O. Moon, K. E. Thomas, V. M. Ferreira, S. E. Petersen, S. Neubauer, and S. K. Piechnik, "Quality Control-Driven Image Segmentation Towards Reliable Automatic Image Analysis in Large-Scale Cardiovascular Magnetic Resonance Aortic Cine Imaging," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds.), vol. 11765 LNCS, (Cham), pp. 750–758, Springer International Publishing, 2019.

[49] E. Hann, I. A. Popescu, Q. Zhang, R. A. Gonzales, A. Barutçu, S. Neubauer, V. M. Ferreira, and S. K. Piechnik, "Deep Neural Network Ensemble for On-the-Fly Quality Control-Driven Segmentation of Cardiac MRI T1 Mapping," *Medical Image Analysis*, p. 102029, mar 2021.

[50] M. F. O'Rourke, M. E. Safar, and V. Dzau, "The Cardiovascular Continuum extended: Aging effects on the aorta and microvasculature," *Vascular Medicine*, vol. 15, pp. 461–468, dec 2010.

[51] A. Redheuil, W. C. Yu, C. O. Wu, E. Mousseaux, A. De Cesare, R. Yan, N. Kachenoura, D. Bluemke, and J. A. Lima, "Reduced ascending aortic strain and distensibility: Earliest manifestations of vascular aging in humans," *Hypertension*, vol. 55, pp. 319–326, feb 2010.

[52] A. Redheuil, C. O. Wu, N. Kachenoura, Y. Ohyama, R. T. Yan, A. G. Bertoni, G. W. Hundley, D. A. Duprez, D. R. Jacobs, L. B. Daniels, C. Darwin, C. Sibley, D. A. Bluemke, J. A. C. Lima, and J. A. Lima, "Proximal aortic distensibility is an independent predictor of all-cause mortality and incident CV events: the MESA study.," *Journal of the American College of Cardiology*, vol. 64, pp. 2619–2629, dec 2014.

[53] M. Sanghvi, L. Biasiolli, N. Aung, J. A. Cooper, K. Fung, E. Lukaschuk, J. M. Paiva, V. Carapella, E. Hann, J. J. Rayner, K. Werys, H. Puchta, S. K. Piechnik, S. Neubauer, and S. E. Petersen, "The impact of modifiable cardiovascular risk factors on aortic distensibility: insights from the UK Biobank — European Heart Journal - Cardiovascular Imaging — Oxford Academic," in *European Heart Journal - Cardiovascular Imaging*, vol. 20, 2019.

[54] K. Fung, L. Biasiolli, E. Hann, N. Aung, J. Paiva, E. Lukaschuk, M. Sanghvi, V. Carapella, J. Rayner, K. Werys, H. Puchta, K. Thomas, N. Moon, M. Khanji, S. Neubauer, S. Piechnik, P. B. Munroe, and S. Petersen, "Effect of coffee consumption on arterial stiffness from UK biobank imaging study," in *Heart*, vol. 105, pp. A8.2–A10, BMJ, may 2019.

[55] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321–331, jan 1988.

[56] X. Li, B. Aldridge, R. Fisher, and J. Rees, "Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation," in *Proceedings - International Symposium on Biomedical Imaging*, pp. 1438–1441, IEEE, mar 2011.

[57] J. C. Moon, D. R. Messroghli, P. Kellman, S. K. Piechnik, M. D. Robson, M. Ugander, P. D. Gatehouse, A. E. Arai, M. G. Friedrich, S. Neubauer, J. Schulz-Menger, and E. B. Schelbert, "Myocardial T1 mapping and extracellular volume quantification: A Society for Cardiovascular Magnetic Resonance

(SCMR) and CMR Working Group of the European Society of Cardiology consensus statement," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, p. 92, oct 2013.

[58] D. R. Messroghli, J. C. Moon, V. M. Ferreira, L. Grosse-Wortmann, T. He, P. Kellman, J. Mascherbauer, R. Nezafat, M. Salerno, E. B. Schelbert, A. J. Taylor, R. Thompson, M. Ugander, R. B. Van Heeswijk, and M. G. Friedrich, "Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2 and extracellular volume: A consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imagin," *Journal of Cardiovascular Magnetic Resonance*, vol. 19, p. 75, dec 2017.

[59] J. Čelutkienė, C. M. Plymen, F. A. Flachskampf, R. A. de Boer, J. Grapsa, R. Manka, L. Anderson, M. Garbi, V. Barberis, P. P. Filardi, P. Gargiulo, J. L. Zamorano, M. Lainscak, P. Seferovic, F. Ruschitzka, G. M. Rosano, and P. Nihoyannopoulos, "Innovative imaging methods in heart failure: a shifting paradigm in cardiac assessment. Position statement on behalf of the Heart Failure Association of the European Society of Cardiology," *European Journal of Heart Failure*, vol. 20, pp. 1615–1633, dec 2018.

[60] H. H. Huang, C. Y. Huang, C. N. Chen, Y. W. Wang, and T. Y. Huang, "Automatic regional analysis of myocardial native T1 values: left ventricle segmentation and AHA parcellations," *International Journal of Cardiovascular Imaging*, vol. 34, pp. 131–140, jan 2018.

[61] S. K. Piechnik, V. M. Ferreira, E. Dall'Armellina, L. E. Cochlin, A. Greiser, S. Neubauer, and M. D. Robson, "Shortened Modified Look-Locker Inversion recovery (ShMOLLI) for clinical myocardial T1-mapping at 1.5 and 3 T within a 9 heartbeat breathhold," *Journal of Cardiovascular Magnetic Resonance*, vol. 12, p. 69, dec 2010.

[62] E. Dall'Armellina, S. K. Piechnik, V. M. Ferreira, Q. L. Si, M. D. Robson, J. M. Francis, F. Cuculi, R. K. Kharbanda, A. P. Banning, R. P. Choudhury, T. D. Karamitsos, and S. Neubauer, "Cardiovascular magnetic resonance by non contrast T1-mapping allows assessment of severity of injury in acute myocardial infarction," *Journal of Cardiovascular Magnetic Resonance*, vol. 14, p. 15, feb 2012.

[63] V. M. Ferreira, S. K. Piechnik, E. Dallarmellina, T. D. Karamitsos, J. M. Francis, R. P. Choudhury, M. G. Friedrich, M. D. Robson, and S. Neubauer, "Non-contrast T1-mapping detects acute myocardial edema with high diagnostic accuracy: A comparison to T2-weighted cardiovascular magnetic resonance," *Journal of Cardiovascular Magnetic Resonance*, vol. 14, p. 42, jun 2012.

[64] S. Dass, J. J. Suttie, S. K. Piechnik, V. M. Ferreira, C. J. Holloway, R. Banerjee, M. Mahmod, L. Cochlin, T. D. Karamitsos, M. D. Robson, H. Watkins, and S. Neubauer, "Myocardial tissue characterization using magnetic resonance non-contrast T1 mapping in hypertrophic and dilated cardiomyopathy," *Circulation: Cardiovascular Imaging*, vol. 5, pp. 726–733, nov 2012.

[65] S. K. Piechnik, V. M. Ferreira, A. J. Lewandowski, N. A. Ntusi, R. Banerjee, C. Holloway, M. B. Hofman, D. M. Sado, V. Maestrini, S. K. White, M. Lazdam, T. Karamitsos, J. C. Moon, S. Neubauer, P. Leeson, and M. D. Robson, "Normal variation of magnetic resonance T1 relaxation times in the human population at 1.5 T using ShMOLLI," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, p. 13, jan 2013.

[66] S. Bull, S. K. White, S. K. Piechnik, A. S. Flett, V. M. Ferreira, M. Loudon, J. M. Francis, T. D. Karamitsos, B. D. Prendergast, M. D. Robson, S. Neubauer, J. C. Moon, and S. G. Myerson, "Human non-contrast T1 values and correlation with histology in diffuse fibrosis," *Heart*, vol. 99, pp. 932–937, jul 2013.

[67] T. D. Karamitsos, S. K. Piechnik, S. M. Banypersad, M. Fontana, N. B. Ntusi, V. M. Ferreira, C. J. Whelan, S. G. Myerson, M. D. Robson, P. N. Hawkins,

S. Neubauer, and J. C. Moon, "Noncontrast T1 mapping for the diagnosis of cardiac amyloidosis," *JACC: Cardiovascular Imaging*, vol. 6, pp. 488–497, apr 2013.

[68] V. M. Ferreira, S. K. Piechnik, E. Dall'Armellina, T. D. Karamitsos, J. M. Francis, N. Ntusi, C. Holloway, R. P. Choudhury, A. Kardos, M. D. Robson, M. G. Friedrich, and S. Neubauer, "T1 Mapping for the diagnosis of acute myocarditis using CMR: Comparison to T2-Weighted and late gadolinium enhanced imaging," *JACC: Cardiovascular Imaging*, vol. 6, pp. 1048–1058, oct 2013.

[69] V. M. Ferreira, S. K. Piechnik, M. D. Robson, S. Neubauer, and T. D. Karamitsos, "Myocardial tissue characterization by magnetic resonance imaging: Novel applications of T1 and T2 mapping," *Journal of Thoracic Imaging*, vol. 29, pp. 147–154, may 2014.

[70] N. A. Ntusi, S. K. Piechnik, J. M. Francis, V. M. Ferreira, A. B. Rai, P. M. Matthews, M. D. Robson, J. Moon, P. B. Wordsworth, S. Neubauer, and T. D. Karamitsos, "Subclinical myocardial inflammation and diffuse fibrosis are common in systemic sclerosis - A clinical study using myocardial T1-mapping and extracellular volume quantification," *Journal of Cardiovascular Magnetic Resonance*, vol. 16, p. 21, mar 2014.

[71] V. M. Ferreira, S. K. Piechnik, E. Dall'Armellina, T. D. Karamitsos, J. M. Francis, N. Ntusi, C. Holloway, R. P. Choudhury, A. Kardos, M. D. Robson, M. G. Friedrich, and S. Neubauer, "Native T1-mapping detects the location, extent and patterns of acute myocarditis without the need for gadolinium contrast agents," *Journal of Cardiovascular Magnetic Resonance*, vol. 16, p. 36, may 2014.

[72] M. Mahmod, S. K. Piechnik, E. Levelt, V. M. Ferreira, J. M. Francis, A. Lewis, N. Pal, S. Dass, H. Ashrafian, S. Neubauer, and T. D. Karamitsos, "Adenosine stress native T1 mapping in severe aortic stenosis: evidence for a role of the intravascular compartment on myocardial T1 values," *Journal of cardiovascular*

magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance, vol. 16, p. 92, dec 2014.

[73] N. A. Ntusi, S. K. Piechnik, J. M. Francis, V. M. Ferreira, P. M. Matthews, M. D. Robson, P. B. Wordsworth, S. Neubauer, and T. D. Karamitsos, "Diffuse myocardial fibrosis and inflammation in rheumatoid arthritis: Insights from CMR T1 Mapping," *JACC: Cardiovascular Imaging*, vol. 8, pp. 526–536, may 2015.

[74] E. Levelt, M. Mahmod, S. K. Piechnik, R. Ariga, J. M. Francis, C. T. Rodgers, W. T. Clarke, N. Sabharwal, J. E. Schneider, T. D. Karamitsos, K. Clarke, O. J. Rider, and S. Neubauer, "Relationship between left ventricular structural and metabolic remodeling in type 2 diabetes," *Diabetes*, vol. 65, pp. 44–52, oct 2016.

[75] V. M. Ferreira, R. S. Wijesurendra, A. Liu, A. Greiser, B. Casadei, M. D. Robson, S. Neubauer, and S. K. Piechnik, "Systolic ShMOLLI myocardial T1-mapping for improved robustness to partial-volume effects and applications in tachyarrhythmias," in *Journal of Cardiovascular Magnetic Resonance*, vol. 17, p. 77, BioMed Central, dec 2015.

[76] N. Ntusi, E. O'Dwyer, L. Dorrell, E. Wainwright, S. Piechnik, G. Clutton, G. Hancock, V. Ferreira, P. Cox, M. Badri, T. Karamitsos, S. Emmanuel, K. Clarke, S. Neubauer, and C. Holloway, "HIV-1-Related Cardiovascular Disease Is Associated with Chronic Inflammation, Frequent Pericardial Effusions, and Probable Myocardial Edema," *Circulation: Cardiovascular Imaging*, vol. 9, p. e004430, mar 2016.

[77] V. M. Ferreira, M. Marcelino, S. K. Piechnik, C. Marini, T. D. Karamitsos, N. A. Ntusi, J. M. Francis, M. D. Robson, J. R. Arnold, R. Mihai, J. D. Thomas, M. Herincs, Z. K. Hassan-Smith, A. Greiser, W. Arlt, M. Korbonits, N. Karavitaki, A. B. Grossman, J. A. Wass, and S. Neubauer, "Pheochromocytoma is

characterized by catecholamine-mediated myocarditis, focal and diffuse myocardial fibrosis, and myocardial dysfunction," *Journal of the American College of Cardiology*, vol. 67, pp. 2364–2374, may 2016.

[78] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," dec 2014.

[79] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, pp. 205–219, aug 2015.

[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[81] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift," 2019.

[82] J. Thagaard, S. Hauberg, B. V. D. Vegt, T. Ebstrup, J. D. Hansen, and A. B. Dahl, "Can you trust predictive uncertainty under real dataset shifts in digital pathology ?," in *Medical Image Computing and Computer Assisted Intervention*, pp. 1–10, 2020.

[83] D. R. Messroghli, A. Radjenovic, S. Kozerke, D. M. Higgins, M. U. Sivananthan, and J. P. Ridgway, "Modified Look-Locker inversion recovery (MOLLI) for high-resolutionT1 mapping of the heart," *Magnetic Resonance in Medicine*, vol. 52, pp. 141–146, jul 2004.

[84] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable," *IEEE Transactions on Medical Imaging*, vol. 31, pp. 153–163, feb 2012.

[85] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, "Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations," *arXiv*, jun 2019.

[86] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 384–395, may 2018.