MyoPS A Benchmark of Myocardial Pathology Segmentation Combining Three-Sequence Cardiac Magnetic Resonance Images

Citation for published version:
Li, L, Wu, F, Wang, S, Luo, X, Martin-Isla, C, Zhai, S, Zhang, J, Liu7, Y, Zhang, Z, Ankenbrand, MJ, Jiang, H, Zhang, X, Wang, L, Arega, TW, Altunok, E, Zhao, Z, Li, F, Ma, J, Yang, X, Puybareau, E, Oksuz, I, Bricq, S, Li, W, Punithakumar, K, Tsaftaris, SA, Schreiber, LM, Yang, M, Liu, G, Xia, Y, Wang, G, Escalera, S & Zhuang, X 2023, 'MyoPS A Benchmark of Myocardial Pathology Segmentation Combining Three-Sequence Cardiac Magnetic Resonance Images', *Medical Image Analysis*, vol. 87, 102808. https://doi.org/10.1016/j.media.2023.102808

Digital Object Identifier (DOI):
10.1016/j.media.2023.102808

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Medical Image Analysis

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

MyoPS: A Benchmark of Myocardial Pathology Segmentation Combining Three-Sequence Cardiac Magnetic Resonance Images

Lei Li[1,2†], Fuping Wu[1†], Sihan Wang[1†], Xinzhe Luo[1], Carlos Martín-Isla[3], Shuwei Zhai[5], Jianpeng Zhang[6], Yanfei Liu[7], Zhen Zhang[9], Markus J. Ankenbrand[10], Haochuan Jiang[11,12], Xiaoran Zhang[13], Linhong Wang[15], Tewodros Weldebirhan Arega[16], Elif Altunok[17], Zhou Zhao[18], Feiyan Li[15], Jun Ma[19], Xiaoping Yang[20], Elodie Puybareau[18], Ilkay Oksuz[17], Stephanie Bricq[16], Weisheng Li[15], Kumaradevan Punithakumar[14], Sotirios A. Tsaftaris[11], Laura M. Schreiber[10], Mingjing Yang[9], Guocai Liu[7,8], Yong Xia[6], Guotai Wang[5], Sergio Escalera[3,4], Xiahai Zhuang[1*]

[1] *School of Data Science, Fudan University, Shanghai, China*
[2] *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China*
[3] *Departament de Matemàtiques & Informàtica, Universitat de Barcelona, Barcelona, Spain*
[4] *Computer Vision Center, Universitat Autònoma de Barcelona, Spain*
[5] *School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[6] *School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China*
[7] *College of Electrical and Information Engineering, Hunan University, Changsha, China*
[8] *National Engineering Laboratory for Robot Visual Perception and Control Technology, Changsha, China*
[9] *College of Physics and Information Engineering, Fuzhou University, Fuzhou, China*
[10] *Chair of Molecular and Cellular Imaging, Comprehensive Heart Failure Center, Wuerzburg University Hospitals, Wuerzburg, Germany*
[11] *School of Engineering, University of Edinburgh, Edinburgh, UK*
[12] *School of Robotics, Xi'an Jiaotong-Liverpool University, Suzhou, China*
[13] *Department of Electrical and Computer Engineering, University of California, Los Angeles, USA*
[14] *Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton, Canada*
[15] *Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecomm-unications, Chongqing, China*
[16] *ImViA Laboratory, Université Bourgogne Franche-Comté, Dijon, France*
[17] *Computer Engineering Department, Istanbul Technical University, Istanbul, Turkey*
[18] *EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France*
[19] *Department of Mathematics, Nanjing University of Science and Technology, Nanjing, China*
[20] *Department of Mathematics, Nanjing University, Nanjing, China*

ARTICLE INFO

Article history:

Myocardial Pathology Segmentation, Multi-Source Images, Cardiac Magnetic Resonance, Multi--Sequence MRI, Benchmark

ABSTRACT

Assessment of myocardial viability is essential in diagnosis and treatment management of patients suffering from myocardial infarction, and classification of pathology on myocardium is the key to this assessment. This work defines a new task of medical image analysis, i.e., to perform myocardial pathology segmentation (MyoPS) combining three-sequence cardiac magnetic resonance (CMR) images, which was first proposed in the MyoPS challenge, in conjunction with MICCAI 2020. Note that MyoPS refers to both myocardial pathology segmentation and the challenge in this paper. The challenge provided 45 paired and pre-aligned CMR images, allowing algorithms to combine the complementary information from the three CMR sequences for pathology segmentation. In this article, we provide details of the challenge, survey the works from fifteen participants and interpret their methods according to five aspects, i.e., preprocessing, data augmentation, learning strategy, model architecture and post-processing. In addition, we analyze the results with respect to different factors, in order to

[*] Senior and corresponding author:Xiahai Zhuang; [†] These authors contribute equally and are the co-first authors: Lei Li (lilei.sky@sjtu.edu.cn), Fuping Wu (17110690006@fudan.edu.cn) and Sihan Wang (shwang21@m.fudan.edu.cn).

examine the key obstacles and explore the potential of solutions, as well as to provide a benchmark for future research. The average Dice scores of submitted algorithms were 0.614 ± 0.231 and 0.644 ± 0.153 for myocardial scars and edema, respectively. We conclude that while promising results have been reported, the research is still in the early stage, and more in-depth exploration is needed before a successful application to the clinics. MyoPS data and evaluation tool continue to be publicly available upon registration via its homepage (www.sdspeople.fudan.edu.cn/zhuangxiahai/0/myops20/).

1. Introduction

1.1. Clinical background

Myocardial infarction (MI) is a major cause of mortality and disability worldwide (Thygesen et al., 2008). Assessment of myocardial viability is essential in the diagnosis and treatment management for patients suffering from MI. In particular, the position and distribution of myocardial infarct (also known as “scar”) and edema could provide important information for selection of patients and delivery of therapies of MI. Specifically, myocardial scars refer to the area where the left ventricle loses viability and is a prominent cause of serious complications, such as heart failure and ventricular arrhythmias (Delgado et al., 2011). Edema is induced by ischemia and reperfusion, and its size reflects the area of ischemic injury in early acute (per-acute) MI (Ruder et al., 2013). The presence of myocardial edema may be associated with a higher hazard of cardiovascular event or death (Friedrich, 2010; Raman et al., 2010). Hence, characterizing the evolution of myocardial scars and edema has important prognostic value and can be used to evaluate the efficacy of future therapies.

Cardiac magnetic resonance (CMR) imaging can be used to determine the effects of acute MI in vivo, as Figure 1 shows. For example, the balanced steady-state free precession (bSSFP) sequence can be used to analyze the left ventricular (LV) volume and wall thickness, as it provides a clear LV boundary. Late gadolinium enhancement (LGE) CMR imaging can visualize infarction, while T2-weighted CMR can depict myocardial edema referring to the area at risk after acute MI. To accurately differentiate nonviable infarct myocardium from viable peri-infarct tissues, Kidambi et al. (2013) defined infarct zone on the 90-day LGE images and peri-infarct zone on the 2-day T2-weighted images acquired from the same patient. Therefore, the edema can be divided into two regions of interest around the infarction: the infarct zone and peri-infarct zone, as Figure 2 (a) shows. The task of our challenge is to segment myocardial pathology by combining the three-sequence CMR images from the same patient, assuming the three sequences are aligned prior to pathology segmentation. This task is illustrated in Figure 1.

1.2. Challenge

As manual segmentation is time-consuming and subjective, automatic myocardial pathology segmentation (MyoPS) is highly demanded. However, automating this segmentation remains challenging, due to the large shape variability of myocardium, indistinguishable boundaries, and the possible poor image quality. Particularly, there are three challenges for the automatic multi-image-based pathology segmentation. Firstly, the intensity distribution of the pathological myocardium in LGE and T2 CMR images is heterogeneous. Secondly, the enhancements of pathologies can be highly variable and complex. The location, shape and size of infarcts and edemas vary greatly across different patients. Finally, the misalignment of inter-sequence images introduces new challenges to combine them for the pathology segmentation.

To the best of our knowledge, few works have been reported for MyoPS combining multi-sequence CMR images (Li et al., 2022a). Most works only segments single pathology, i.e., scars or edema, based on a single CMR sequence. This could be due to the difficulty of correcting the misalignment among different sequences. Therefore, we defined the task of MyoPS where three-sequence CMR images from the same subject were pre-aligned in the challenge event. This was to mitigate the difficulty of misalignment and data missing (Zhuang, 2019), and to encourage the participants to solely focus on the algorithms of MyoPS.

1.3. Motivation

We therefore organized the MyoPS challenge 2020 in conjunction with MICCAI 2020. Specifically, the challenge provided three-sequence CMR from 45 subjects and was aimed to encourage the development of new segmentation algorithms which could combine the complementary information from the three CMR sequences. Twenty-three submissions were evaluated before the deadline, and fifteen teams presented their work at the conference event. In this paper, we introduce the related information, review the methodologies, and analyze their results in detail. Our target is to raise interest in studies on pathology segmentation of myocardium combining multi-source images, which

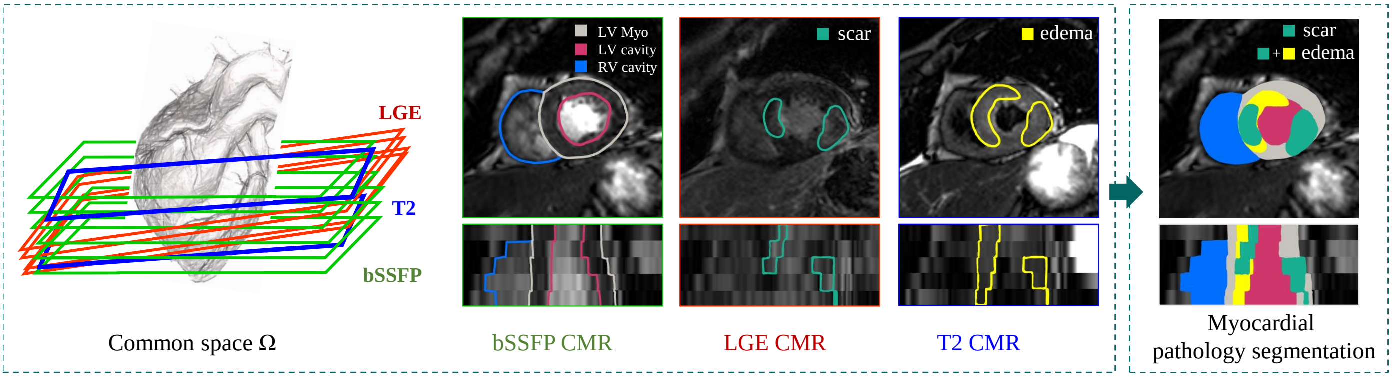


Figure 1. Visualization of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images acquired from the same patient (figure designed referring to Zhuang (2019)).

has been employed for studies of other organs, such as segmentation of brain tumor (Du et al., 2016) and prostate cancer (Vall and Lemaitre, 2016).

The rest of this paper is structured as follows: Section 2 presents an overview of related work in the previous challenges of MICCAI and their benchmarks, as well as the difficulties and current solutions. Section 3 provides details of the materials and evaluation framework from the challenge. Section 4 summarizes the current methods for MyoPS. Section 5 describes the results, followed by discussions in Section 6. Finally, we conclude this work in Section 7.

2. Related work

2.1. Related challenges and benchmarks

In recent years, there are many challenges of computational modeling, segmentation and computer-aided diagnosis for cardiovascular problems. Thanks to those challenges, researchers can develop, test and compare computational algorithms on the same dataset. Table 1 presents the recent challenges and public datasets for cardiac segmentation. One can see that only Karim et al. (2016) and Karim et al. (2013) focused on the LV/ left atrial (LA) scar segmentation from LGE CMR. None of them was aimed to combine multi-source images. Though there was a challenge with multi-sequence CMR images, it was aimed to segment myocardium from LGE CMR by referring to the training images from other sequences (Zhuang et al., 2022). In contrast, MyoPS challenge was aimed to segment and classify the pathology of myocardium combining the complementary information related to the pathology and morphology from the three-sequence CMR, i.e., the bSSFP, T2 and LGE CMR.

Current pathology segmentation challenges combining multi-modality images mainly came from brain image analysis. For example, multi-modal brain tumor segmentation (BraTS) challenge was organized in conjunction with the MICCAI 2012-2020 conferences. The BraTS challenge provided native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 fluid attenuated inversion recovery (T2-FLAIR) MR images for the brain tumor segmentation. The first corresponding benchmark study (Menze et al., 2014) summarized eleven submitted algorithms, which were all conventional methods, such as fuzzy clustering, level set, and support vector machine. It found that different algorithms could achieve high performance on a specific subregion, but no one performed consistently better than the others for all subregions. The next benchmark study (Bakas et al., 2018) of the challenge aimed to assess the state-of-the-art machine learning methods for multimodal brain tumor segmentation, during BraTS challenge 2012-2018. Ischemic stroke lesion segmentation (ISLES) challenge offered at least the set of T1-weighted, T2, diffusion weighted imaging (DWI) and FLAIR MR sequences for each case with a diagnosis of ischemic stroke. Their benchmark study found that no algorithmic characteristic of any methods was proved better than others, and emphasized the importance of the characteristics of stroke lesion appearances, their evolution and the observed challenges (Maier et al., 2017).

Other related challenges for other organs include the I2CVB (Vall and Lemaitre, 2016) and CHAOS (Kavur et al., 2021). I2CVB provided a multi-parametric MR image dataset, including T2 MR, dynamic contrast enhanced (DCE) MR, DWI MR and MR spectroscopic imaging data, and was aimed for prostate cancer segmentation (Vall and Lemaitre, 2016). CHAOS combined CT and MR images from the abdomen for organ segmentation, including liver, kidneys and spleen (Kavur et al., 2021). To the best of our knowledge, there is still no challenge/ public available dataset on cardiac pathology segmentation combining multi-source images.

2.2. State-of-the-art myocardial pathology segmentation

A short overview of previously published algorithms related to MyoPS is presented here, and summarized in Table 2. One can see that only Baron et al. (2008) segmented both myocardial scar and edema, respectively from LGE CMR and T2 CMR. Other studies only focus on one of them. In specific, for scar segmentation the most widespread methods are mainly based on thresholding, such as n-SD and full-width-at-half-maximum (FWHM) (Karim et al., 2016; Sandfort et al., 2017). It is mostly

Table 1. Summary of previous challenges related to the cardiac segmentation from MICCAI/ ISBI society. LV: left ventricle; Myo: myocardium; RV: right ventricle; LA: left atrium; WHBP: whole heart blood pool; WH: whole heart; SM: single modality; MM: multi-modality; MI: myocardial infarction; MH: myocardial hypertrophy; ConHD: congenital heart disease; DCM: dilated cardiomyopathy; CorHD: coronary heart disease; AF: atrial fibrillation; HCM: Hypertrophic cardiomyopathy; HHD: Hypertensive Heart Disease; ARV: abnormal right ventricle; AHS: athlete's heart syndrome; IHD: ischemic heart disease; LVNC: left ventricle non-compaction; ‡: multi-center datasets.

Challenge	Year	Source	Data info	Target	Pathologies
Radau et al. (2009)	2009	SM	45 bSSFP CMR	LV, Myo	MI, MH
Suinesiaputra et al. (2011)	2011	MM	200 bSSFP CMR	LV, Myo	MI
Petitjean et al. (2015)	2012	SM	48 bSSFP CMR	RV	ConHD
Karim et al. (2016)	2012	SM	30 LGE CMR	LV scars	MI
Karim et al. (2013)	2013	SM‡	60 LGE CMR	LA scar	AF
Tobon-Gomez et al. (2015)	2013	MM	30 CT, 30 bSSFP CMR	LA	AF
Karim et al. (2018)	2016	MM	10 CT, 10 black-blood CMR	LA wall	AF
Moghari et al. (2016)	2016	SM	20 bSSFP CMR	WHBP, Myo	ConHD
Bernard et al. (2018)	2017	SM	150 bSSFP CMR	LV, Myo, RV	MI, MH, DCM, abnormal RV
Zhuang et al. (2019)	2017	MM‡	60 CT, 60 bSSFP CMR	WH	AF, ConHD, CorHD
Xiong et al. (2020)	2018	MM	150 LGE CMR	LA	AF
Zhuang et al. (2022)	2019	MM	45 bSSFP, LGE, T2 CMR	LV, Myo, RV	MI
Lalande et al. (2020)	2020	SM	150 LGE	LV scars	MI
Campello et al. (2021)	2020	SM‡	150 bSSFP CMR	RV, LV, Myo	HCM, DCM, HHD, ARV, AHS, and IHD

Table 2. Summary of current myocardial pathology segmentation algorithms. CF: connectivity filtering; RG: region growing; Error: error in predicted scar/edema percentage; RF: random forest; RMSE: root mean squared error; HD: Hausdorff distance; LCE: late contrast enhancement; ICC: intraclass correlation coefficient; GMM: Gaussian mixture model.

Reference	**Data**	**Target(s)**	**Method**	**Results**
Baron et al. (2008)	22 LGE CMR + T2 CMR	Scar + Edema	Fuzzy clustering	Volume correlation: r > 0.8

Tao et al. (2010)	20 LGE CMR	Scar	Otsu + CF and RG	Dice = 0.83 ± 0.07 & 0.79 ± 0.08;
				Error = 0.0 ± 1.9% & 3.8 ± 4.7%

Lu et al. (2012)	9 LGE CMR	Scar	Graph-cuts	N/A
Sandfort et al. (2017)	34 LCE CT	Scar	Adaptive threshold	Dice = 0.47; ICC (volume/area) = 0.96/0.87

Xu et al. (2018)	165 cine CMR	Scar	LSTM-RNN + optical flow	Accuracy = 0.95; Kappa = 0.91;
				Dice = 0.90; RMSE = 0.72 mm;
				HD = 5.91 mm

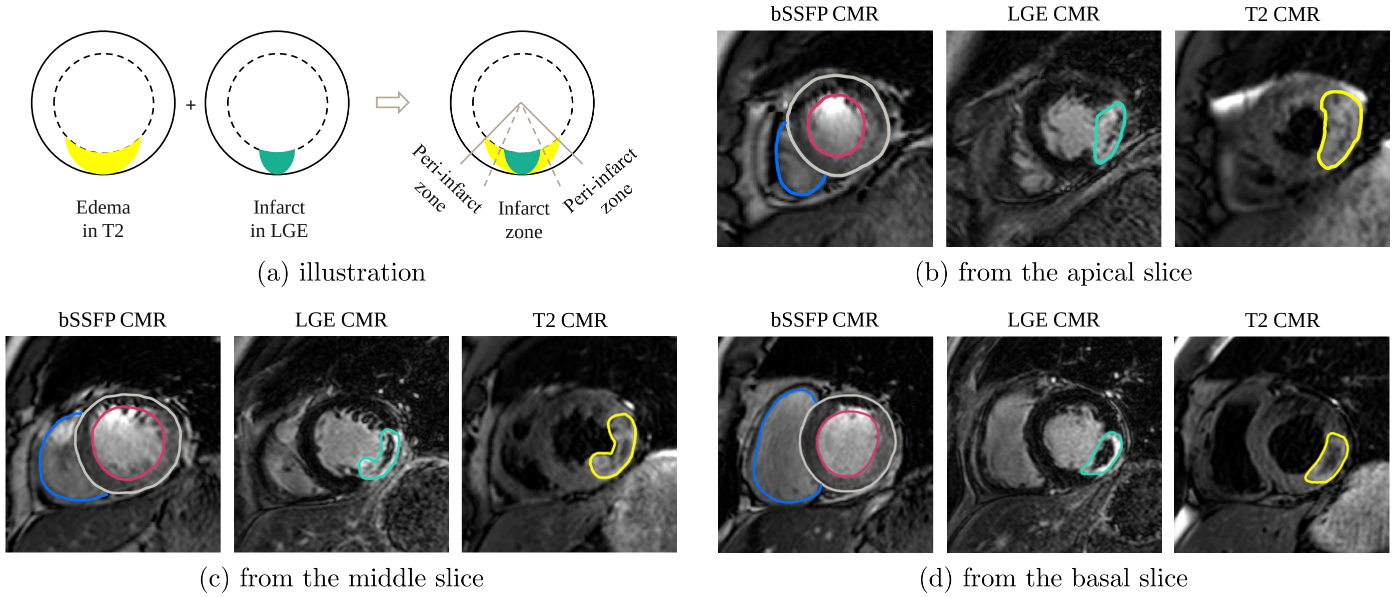
Kurzendorfer et al. (2018)	30 LGE CMR	Scar	Fractal analysis + RF	Dice = 0.64 ± 0.17
Moccia et al. (2019)	30 LGE CMR	Scar	FCN	Sensitivity = 0.881; Dice = 0.713

Zabihollahy et al. (2019)	34 LGE CMR	Scar	CNN	Dice = 0.936 ± 0.026;
				Jaccard = 0.881 ± 0.470

Kadir et al. (2011)	16 T2 CMR	Edema	Morphological filtering + threshold	Volume correlation: r > 0.8;
				Error = 9.95± 3.90%

Gao et al. (2013)	25 T2 CMR	Edema	Rayleigh-GMM	Dice = 0.74

attributed to the relatively evident intensity contrast between scarring areas and background inside the myocardium. Instead of simply using thresholding, Tao et al. (2010) combined it with connectivity filtering and region growing. Other conventional methods were also employed, such as fuzzy clustering (Baron et al., 2008), graph-cuts (Lu et al., 2012), and fractal analysis with random forest (Kurzendorfer et al., 2018). Recently, thanks to the great advance in deep learning (DL), Moccia et al. (2019) and Zabihollahy et al. (2019) employed fully convolutional networks (FCN) and convolutional neural networks (CNN) for LV scar segmentation. One can see that most works extracted scars from LGE CMR or late contrast enhancement CT, where scars are enhanced to distinguish them from non-scarring areas of the myocardium. However, for the patients with chronic end-stage kidney diseases, the administration of gadolinium contrast agent is dangerous. Therefore, Xu et al. (2018) proposed an effective method to directly obtain the position, shape, and size of an infarction area from a raw CMR sequence, i.e., cine CMR. Compared with LV scar segmentation, there were few works on LV edema segmentation. The two works listed in Table 2 for edema segmentation were both based on conventional segmentation methods, and no DL-based method was reported, to the best of our knowledge. Note that here we only focus on the LV myocardial pathology, and for the literature of



(a) illustration (b) from the apical slice (c) from the middle slice (d) from the basal slice

Figure 2. (a) Illustration of the relationship between myocardial infarct (also referred to as scars) and edema; (b)-(d) exemplary three-sequence CMR images superimposed with contours of gold standard segmentation at different slice positions.

another similar topic (i.e., LA myocardial pathology) one could refer to the review (Li et al., 2022b).

3. Materials and setup

3.1. Data

MyoPS challenge provides 45 paired three-sequence CMR images (bSSFP, LGE and T2 CMR) acquired from the same patient. The three CMR sequences were all breath-hold, multi-slice, acquired from the cardiac short-axis views using Philips Achieva 1.5T. All patients are male with acute MI, and the average age and weight are 56.2 ± 7.92 years and 74.4±5.65 kg, respectively. Table 3 provides the acquisition parameters of the three CMR sequences. The data acquisition had been anonymized, and approved by the institutional ethics board. All the data have been preprocessed using the MvMM method (Zhuang, 2019), to align the three-sequence images of the same patient into a common space and to resample them into the same spatial resolution. Each sequence typically contains 2–6 slices with an in-plane resolution of 0.73-0.76 × 0.73-0.76 mm and image size ranging from 412 × 408 to 512 × 515. Figure 2 (b)-(d) provide the exemplary images of the three sequences, with contours of gold standard segmentation results superimposed on, from respectively the apical, middle and basal slices.

For generating the gold standard segmentation, three observers were employed to manually label the LV blood pool, right ventricular (RV) blood pool, and LV myocardium (Myo) from all the three CMR sequences. In addition, LV myocardial scar and edema were manually delineated from LGE and T2 CMR images, respectively. The observers were well-trained raters who were post-graduate students either in biomedical engineering or medical imaging field. They followed the provided instructions: (1) The position of myocardial scars and edema must be located inside the myocardium, which includes the papillary muscle. (2) Cine CMR image was utilized as a reference to delimit the myocardial and LV/ RV regions of LGE and T2 CMR. T2 was used to guide the scar segmentation of LGE, and LGE was also employed to guide the edema annotations of T2. (3) The annotation of scars is contained in that of edema. The manual labeling was performed slice-by-slice using a brush tool in the software ITK-SNAP (Yushkevich et al., 2006). All the manual segmentation results were validated by three experts in cardiac anatomy before used in the construction of gold standard segmentation. The final segmentation was obtained by averaging the multiple manual delineations using a shape-based approach (Rohlfing and Maurer, 2006). The inter-observer variations of manual scar and edema segmentation in terms of Dice overlap were 0.569 ± 0.198 and 0.701 ± 0.168, respectively. The manual segmentation of edema was evidently more consistent between the raters than that of scar segmentation, probably due to the fact that the regions of edema are generally larger (in terms of size) and less patchy (in terms of shape) from T2 images.

Finally, for the challenge event we split the data into two sets, i.e., the training set, including validation images and consisting of 25 pairs, and the test set composed of 20 pairs.

3.2. Evaluation metrics

Though the labels of LV, RV and Myo were provided for the training data, the evaluation of test data only focused on the myocardial pathology segmentation, i.e., scars and edema. To evaluate the segmentation accuracy, we cal-

Table 3. Image acquisition parameters of the MyoPS challenge data and image parameter before pre-processing. ED: end-diastolic.

Sequence	Imaging type	TR/ TE (ms)	Slice spacing (thickness + gap)	In-plane resolution
LGE	T1-weighted	3.6/ 1.8	5 mm	0.75×0.75 mm
T2	T2-weighted, black blood	2000/ 90	12-20 mm	1.35×1.35 mm
bSSFP	Cine sequence (ED phase)	2.7/ 1.4	8-13 mm	1.25×1.25 mm

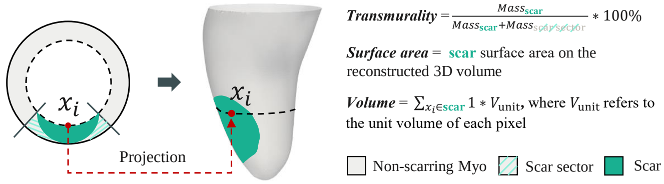

Transmurality = Mass_scar / (Mass_scar + Mass_scar sector) * 100%
Surface area = scar surface area on the reconstructed 3D volume
Volume = ∑_{x_i∈scar} 1 * V_unit, where V_unit refers to the unit volume of each pixel
x_i
x_i
Projection
Non-scarring Myo
Scar sector
Scar

Figure 3. Sketch map of clinical measures employed in MyoPS evaluation. Note that here we only employ the clinical measure calculation for scars as an example.

culated the Dice score of scar and edema segmentation separately,

$$\mathrm{Dice}(V_{\mathrm{seg}}, V_{\mathrm{GD}}) = \frac{2\,|V_{\mathrm{seg}} \cap V_{\mathrm{GD}}|}{|V_{\mathrm{seg}}| + |V_{\mathrm{GD}}|}, \tag{1}$$

where V_{GD} and V_{seg} denote the gold standard and automatic segmentation, respectively. Note that without specific indication, these metrics are generally calculated in a 3D volumes, by considering all slices of the target image at the same time, namely subject-wise evaluation. We compute them for 2D slices only when comparing the performances with regard to different slice positions, namely slice-wise evaluation. Moreover, Dice score of scars is annotated as Dice$^{\ominus}$ when the ground truth contains no scar.

In addition, we employed three statistical measurements, i.e., Accuracy (ACC), Sensitivity (SEN), and Specificity (SPE) of the pathology (positive) and healthy myocardium (negative) classification, which are defined as,

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + FN + TN}, \tag{2}$$

$$\mathrm{SEN} = \frac{TP}{TP + FN}, \tag{3}$$

and

$$\mathrm{SPE} = \frac{TN}{TN + FP}, \tag{4}$$

where TP and FP respectively stand for the number of voxels of true and false positive myocardial pathologies; and TN and FN denote the true and false negatives, respectively. Note that in this task, the participants were required to solely report the segmentation on pathologies, i.e., to output the voxels labeled as scars or edema in the images, thus the remaining voxels of myocardium not classified as pathologies by an algorithm were then considered as TN. Furthermore, there were cases without pathology, for which we define the sensitivity of a segmentation result to be one, i.e., SEN=1.

We also consider three clinical indices, i.e., transmurality, surface area and volume of myocardial pathologies, which could be of important clinical value in MI (Ortiz-Pérez et al., 2007). For example, transmurality could be associated with the severity of ventricular wall motion abnormalities at rest (Schuijf et al., 2004). The transmurality of myocardial pathology was computed for each segment as the ratio (percent) of hyper-enhanced to non-hyper-enhanced areas (Ørn et al., 2007). The average transmurality for each patient was then calculated as the average of the transmurality of all segments with non-zero transmurality. The surface areas of myocardial pathology were calculated based on the 3D reconstruction of the identified myocardial scar/ edema (Tao et al., 2015). As for the volume, it is defined as the product of the number of voxels belonging to scar/ edema and the unit volume of each voxel. The reader is referred to Figure 3 for an illustration of each index.

3.3. The MyoPS challenge

3.3.1. Organization

We submitted a proposal to the MICCAI challenge submission system to apply for our MyoPS challenge. One can access our challenge proposal in the zenodo website. At the same time, we applied for a CMT platform to run this challenge, mainly managing the paper submission. After preparing all the dataset, we scheduled a timetable for the challenge, including the date of data release (1st April, 2020), result/ paper submission (22nd/ 29th July, 2020), associated workshop and result release (4th Oct, 2020). Besides, we designed the task, the distribution of dataset and evaluation metrics. Note that the organizers were not allowed to participate the challenge.

3.3.2. Registration and submission

To access the challenge dataset, researchers were required to sign a data agreement file and return it to the organizers. Before the conference, participants can train their model with the training data, and submit their results to the organizers for evaluation. Each team was allowed to submit their testing results 2 times at most. After the conference, we have released the encrypted ground truth of test data and corresponding evaluation tool to simplify the evaluation step for subsequent participants. Therefore, in principle they can evaluate their models unlimited times.

The participants were encouraged to summarize their methods and results by submitting a paper to the CMT-MyoPS platform. The format should follow the LNCS style according to the main MICCAI conference guidelines,

Table 4. Summary of source code from the participants of MyoPS 2020 challenge.

Team	Code	Reference
UESTC	*https://github.com/HiLab-git/MyoPS2020*	Zhai et al. (2020)
UBA	*https://github.com/cmartin-isla/MYOPs-challenge-StackedBCDUnet*	Martín-Isla et al. (2020)
NPU	*https://github.com/jianpengz/EfficientSeg*	Zhang et al. (2020a)
UHW	*https://github.com/chfc-cmi/miccai2020-myops*	Ankenbrand et al. (2020)
FZU	*https://github.com/kakazxz/myops*	Zhang et al. (2020c)
NJUST	*https://github.com/JunMa11/MyoPS2020*	Ma (2020)
CQUPT I	*https://github.com/fly1995/2020MyoPS-MF-DFA-Net*	Li and Li (2020)
LRDE	*https://github.com/Zhaozhou-lrde/myops2020_code*	Zhao et al. (2020)
CQUPT II	*https://github.com/LynnHg/cmsunet*	Li et al. (2020b)
HNU	*https://github.com/APhun/MyoPS20-HNU*	Liu et al. (2020)
Edin	*https://github.com/falconjhc/MFU-Net*	Jiang et al. (2020)
UBO	*https://github.com/tewodrosweldebirhan/scar_segmentation_myops2020*	Arega and Bricq (2020)
ITU	*https://github.com/altunokelif/MyoPS2020-CMRsegmentation*	Elif and Ilkay (2020)
UOA	*https://github.com/Voldemort108X/myops20*	Zhang et al. (2020b)

but we did not constraint the pages. For the submitted manuscripts, they will be firstly reviewed by the organizers who will ensure the quality of the paper reaches the publication standard. Then, each paper will be reviewed by more than two reviewers. The review procedure will be double-blinded, similar to the MICCAI submissions. Currently, researchers can still download the MyoPS data and evaluation tool via the challenge webpage.

3.3.3. Participants

As an ongoing event, the challenge has received seventy-six requests of registration before the submission of this manuscript, among which sixty-five teams participated the event before the date of the workshop (Oct 4th, 2020). Twenty-three submitted results were evaluated before the submission deadline, and fifteen algorithms were included for this benchmark work. *Note that the team abbreviations in the remaining of this paper refer both to the teams and their corresponding methods, as listed in Table 4.* USTB (Yu et al., 2020) is not listed here as they did not provide open source code of their algorithm.

4. Survey of the methods

For the task of MyoPS, deep learning has attracted the most attention and has also shown great potentials. Similar to other segmentation tasks, the key to success includes the adoption of preprocessing, appropriate architecture of networks and loss function, data augmentation, learning strategy, and post-processing. In this section, we survey the benchmarked methods according to these five aspects. Table 5 and Table 6 summarize the key techniques of them, particularly the latter focuses on architecture and training details of the deep neural networks.

4.1. Preprocessing

Preprocessing can reduce the complexity of data, and facilitate the models to learn the target knowledge without considering the unnecessary variations. The widely adopted techniques include cropping regions of interest (ROI) and intensity normalization.

As the pathology to be segmented exists only in the LV, most of the peripheral areas of the background are in fact redundant. To reduce the complexity from background, all the teams cropped ROIs from the original images prior to the MyoPS. For example, USTB cropped ROIs of 256 × 256 pixels (Yu et al., 2020), and FZU cropped a small ROI and resized into images of 128 × 128 pixels (Zhang et al., 2020c). Another method was to perform a coarse segmentation on the images, to localize the position of LV, and then extract ROIs automatically. For example, UBA adopted a U-Net to predict the myocardial region, and then cropped the smallest bounding box around the myocardium with a small margin of 10 pixels (Martín-Isla et al., 2020). As cine CMR presents clear structures of LV while lacking appearance of pathological regions, they chose this modality as the input of the localization U-Net. Similarly, NJUST used a U-Net to segment the whole LV, and then cropped LV ROI into 112 × 112 from LGE and T2 CMR based on the segmentation results (Ma, 2020).

Intensity normalization aims to transform the intensity ranges of images into the same one. Z-score is a common and simple method, which normalizes the data into zero mean and unit standard deviation; another one is to linearly transform the intensity range of an image into $[0, 1]$, which was used by UBA. More advanced preprocessing involves the application of contrast enhancement to the images. For example, USTB and UHW employed the method of contrast limited adaptive histogram equalization (CLAHE) (Pizer et al., 1987), which is particularly useful for images with low contrast; and UBA further used histogram equalization on the cropped ROIs to enhance the contrast. For a summary of all the teams, one can refer to Table 5 for details.

4.2. Architecture and loss function

The most common architecture in the benchmarked algorithms is U-Net, which extracts multi-scale features and combines them together with a skip connection strategy. For example, UESTC used U-Net for both coarse and

Table 5. Summary of the benchmarked algorithms. EM: equalization matching; HE: histogram equalization; IN: intensity normalization; RGT: random gamma technique; CLAHE: contrast limited adaptive histogram equalization; SA: simple augmentation techniques, including random rotation, random flipping, random scaling, ransom shifting, random cropping, random warping and horizontally flipping.

Team	Pre-processing	Method type	Data augmentation	Post-processing
UESTC	crop	two-stage, weighted ensemble	SA	None
UBA	crop, IN into [0,1], HE	two-stage, ensemble	SA, image synthesis	anatomical constraints, morphological operations
NPU	crop, z-score	end-to-end	SA	remove small isolated regions
USTB	crop, z-score, CLAHE, EM	end-to-end	SA, brightness, contrast shift, non-rigid transformation	remove outliers and disconnected regions
UHW	crop, CLAHE	end-to-end, ensemble	SA, brightness, contrast shift, transformation with simulated MR artifacts	None
FZU	crop	two-stage	SA	None
NJUST	crop, z-score	two-stage	SA, brightness, contrast shift	None
CQUPT I	crop, z-score	end-to-end	SA, contrast adjustment, transpose	None
LRDE	crop, z-score	end-to-end, cascaded	None	keep the largest connected component
CQUPT II	crop, z-score	end-to-end	SA, mirror, reverse	None
HNU	crop, HE, RGT	two-stage	SA	None
Edin	crop	end-to-end	SA	None
UBO	crop, z-score	end-to-end, cascaded	None	connected component analysis, morphological operations
ITU	crop, IN with zero mean	end-to-end	SA, elastic transformation, image dropping out	None
UOA	crop, IN	end-to-end	SA	retain the largest connected component and remove holes

fine segmentation stages. NPU adopted EfficientNets (Tan and Le, 2019) as the encoder to extract features from the CMR sequences. The other useful techniques for feature extraction are the dense connection and attention strategy. For example, UBA employed the BCDU-Net (Azad et al., 2019) to segment the pathologies. BCDU-Net is an extension of U-Net and reuses feature maps via dense connections. USTB embedded a channel attention module and a space attention module at the bottom layer of a U-Net model. The former module can selectively emphasize feature association among different channel maps, and the latter captures the long-range dependencies on feature maps. The effectiveness of these modules was verified in their ablation study. Moreover, FZU extracted features from the three sequences separately. To avoid information redundancy of these features, they adopted the channel attention to emphasize the informative features and suppress useless ones.

As to the selection of loss functions, the most commonly used are Dice loss and cross entropy loss. Nevertheless, boundary loss can also be used to boost the model performance, which is demonstrated in the work of FZU. This could be attributed to its ability to enforce the model to pay more attention to boundary regions. Finally, Table 6 provides a summary of model designing and training for the benchmarked methods.

4.3. Data augmentation

As the shapes of the myocardium and their pathologies have large variations, the training images could be insufficient, leading to the over-fitting problem of deep learning. Data augmentation has proven to be effective in improving the generalization ability of resulting models (Takahashi et al., 2019). We group the augmentation techniques into two categories, i.e., online and offline augmentation.

The online augmentation includes the random rotation, scaling, shifting, flipping, non-rigid transformations, as well as brightness and contrast adjustment. For example, USTB adopted the elastic-transform, grid-distortion and optical-distortion to transform the training images non-rigidly. Experiments showed that this augmentation improved the Dice score by about 8% for scar segmentation (Yu et al., 2020).

The offline augmentation mainly refers to image synthesis. UBA did a comprehensive synthesis operation (Martín-Isla et al., 2020). They utilized the semantic image synthesis with spatially-adaptive normalization (SPADE) method (Park et al., 2019), to achieve style transfer, pathology rotation, epicardial warping and pathology dilation/ erosion. Their ablation study demonstrated that these morphological and style transformations could improve the performance significantly. Interestingly, they found that the style transfer was the most effective, while morphological augmentations, such as the scar and edema dilation and erosion, had limited gains.

Table 6. Network architectures and training details of the benchmarked algorithms. CE: cross entropy; BCE: binary cross entropy; MI: mutual information; SE: Squeeze-and-Excitation. Here, x (*) refers to the number of ensemble models, and Efficient-B1/B2/B3 refer to the EfficientNet with different scales.

Team	Architecture	Ensemble (size)	Batch size	Patch size	Loss function	Optimizer	Learning rate	Device
UESTC	U-Net	x (10)	1	160 × 160	CE and Dice loss	SGD	6e-3 (decay)	NVIDIA GeForce RTX 2080 Ti
UBA	U-Net, BCDU-Net	x (15)	8	256 × 256	weighted BCE and Dice loss	Adam	1e-4	NVIDIA 1080 GPU
NPU	EfficientNet for encoder, BiFPN for decoder	Efficient-B1/B2/B3	64/48/32 for B1/B2/B3	288 × 288	CE, Dice and boundary loss	Adam	1e-4 (decay)	RTX 2080 Ti
USTB	Dual attention U-Net	None	8	256 × 256	Dice loss	SGD	1e-3 (decay)	NVIDIA TITAN RTX
UHW	U-Nets (resnet34 backbone)	x (6)	12	256 × 256	CE and Focal loss	Adam	1e-3 (decay)	NVIDIA Tesla K80
FZU	Channel attention based CNN	None	16	128 × 128	Dice loss	Adam	1e-3	NVIDIA GeForce RTX 2080 Ti
NJUST	2D nnU-Net	x (10)	6	112 × 112	CE and Dice loss	SGD	1e-3	NVIDIA V100
CQUPT I	U-Net and a dense connected path	None	4	256 × 256	Weighted CE and Dice loss	Adam	1e-4 (decay)	NVIDIA Geforce RTX 2080 Ti
LRDE	Cascaded U-Net	None	1	240 × 240	BCE	Adam	1e-4	NVIDIA Quadro P6000 GPU
CQUPT II	Multi-scale U-Net	None	6	256 × 256	CE and Dice loss	Adam	1e-4 (decay)	NVIDIA Geforce RTX 2080 Ti
HNU	U-Net, attention-based M-shaped network	None	20	256 × 256	Focal Dice and MSE loss	Adam	3e-4	NVIDIA TITAN V GPU
Edin	Max-Fusion U-Net	None	4	102 × 102 to 288 × 288 (96+16i,1 $\leq$ $i \leq 12$)	Tversky, focal, and unsupervised reconstruction loss	Adam	1e-4	TitanX
UBO	Densenet with inception and SE block	None	16	350 × 350	Logarithmic Dice and region MI loss	Adam	1e-3	NVIDIA Tesla K80
ITU	Residual U-Net	None	8	256 × 256	Dice loss	Adam	1e-3 (decay)	NVIDIA Quadro RTX 6000
UOA	A linear encoder and decoder, with a network module consisting of U-Net, Mask-RCNN and U-Net++	None	8, 2, 8 for the three components in the network module, respectively	256 × 256	Dice loss for U-Net and U-Net++; classification loss, bounding-box loss and CE loss for Mask-RCNN	Adam	1e-5,1e-3,1e-5 for the three components in the network module, respectively	Tesla P100

4.4. Specification of the learning process

As Table 5 shows, five teams implemented their works in a two-stage manner (coarse-to-fine), by extracting ROIs on the myocardium prior to a fine process of pathology segmentation. The others conducted their models in an end-to-end fashion. In addition, the way of utilizing the extracted ROIs were different, which might explain the discrepancies of their results. For example, after obtaining the mask including the RV, LV and Myo, UESTC cropped the ROI from all the three sequences and concatenated them using this mask. They took the concatenation as an input for the final prediction of pathologies. This strategy can help the segmentation model to take advantage of the extracted knowledge. Their experiments on the validation dataset have shown advantages of this setting, particularly on the edema with more than 2% Dice improvement. Similarly, FZU first learned the mask of LV and Myo, and then did an element-wise multiplication between this mask and the image sequences. LRDE used three U-Nets to obtain the mask covering the LV and RV, the mask of Myo, and the mask of all three structures from cine and T2 CMR. These masks were concatenated with the LGE CMR, and then fed into two U-Nets to predict the mask of scar and edema, respectively. In contrast, UBA solely got the mask of LV, and used it to crop the three sequences. NJUST obtained a mask of LV and Myo from cine CMR, and used it to crop the other two sequences for prediction of pathologies.

4.4.1. Ensemble learning strategy

Given the limited training data in this study, it is difficult to know the best model in advance. Also, models trained with different samples or images from different views could learn diverse knowledge. Therefore, several ensemble learning strategies have been employed to reduce uncertainty in models and increase generalization capabilities in this challenge. For example, UESTC performed a weighted ensemble strategy on the predictions of 2D and 2.5D networks. Their experimental results showed that this ensemble strategy delivered better results. UBA adopted a different strategy by generating a number of datasets with synthesized images, and they trained fifteen models using different training data. As their ablation study demonstrated, this ensemble could capture a greater number of non-trivial unconnected components. Similarly, UHW trained 21 models, but they solely selected the 6 top-performing models for the final aggregation.

4.5. Post-processing

Post-processing can be used to remove redundant small patches and refine or regularize the shape of segmentation results to be more realistic. Among all the benchmarked algorithms, only four conducted post-processing to refine their segmentation results. Specifically, UBA firstly reconstructed the myocardium mask into a ring shape by extracting a skeleton. They calculated the distances of pixels around myocardium to the skeleton, and those with distance less than a threshold were then categorized into edema. For the scar segmentation, 3D components smaller than 100 voxels were excluded. Finally, they did the refinement of the joined edema-scar mask by excluding those 3D components of size smaller than 300 voxels. The experiments showed that with this post-processing, the Dice of scar segmentation was improved by almost 3%. NPU simply removed the small isolated segmentation regions. USTB discarded the pixels outside the target area as well as unreasonable and unconnected pathological components, and further filled them with adjacent category pixels. Experiments showed both could improve the pathology segmentation. UOA employed a post-processing step to solely retain the largest connected component of the predicted LV blood pool and LV epicardium. Besides, they applied an operation to remove holes that appear inside the foreground masks before the linear decoder.

4.6. Summary

So far, we have presented the technical trend in the community of MyoPS ranging from preprocessing to post-processing. Firstly, one can see that to achieve satisfactory accuracy, it is better to simplify the input images as much as possible, such as cutting out redundant information via cropping. Moreover, two-stage frameworks which segment the target structure in a coarse-to-fine manner by localizing the target areas beforehand also have been demonstrated to be useful by UESTC and UBA (Zhai et al., 2020; Martín-Isla et al., 2020). Secondly, data augmentation could be very beneficial to boost the robustness of segmentation models. For example, UBA proposed four types of offline augmentation, each of which was demonstrated to be effective (Martín-Isla et al., 2020). Thirdly, the predicted segmentation map can be refined according to the characteristics of the ROI, such as removing the outliers. Finally, ensemble learning strategies were also effective for both scar and edema segmentation, as they could reduce uncertainty in models. Note that although each technique was demonstrated by the participants to be beneficial to boosting the model performance in their dedicated framework, how to combine these techniques in an optimal manner remains unclear.

5. Results

In this section, we present the results of the evaluated algorithms for comparisons, and then analyze several possible factors that may affect the MyoPS performance.

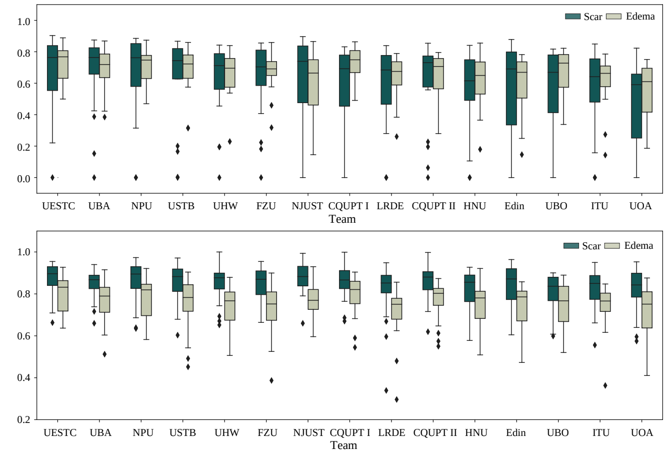


Figure 4. The boxplots of the average Dice and ACC of pathology segmentation obtained by each algorithm.

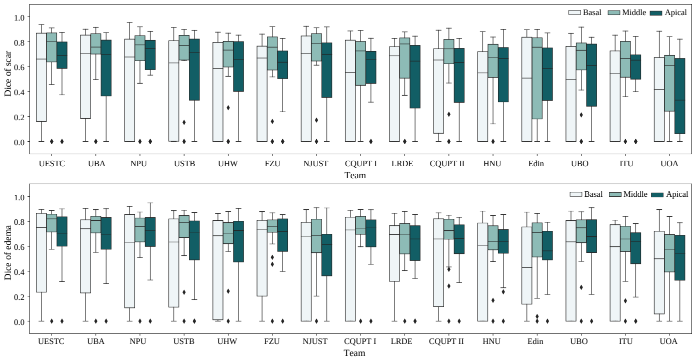


Figure 5. The boxplots of the average Dice of pathology segmentation with respect to different slice positions obtained by each algorithm.

5.1. Overall performance

We present the quantitative results of each team in Table 7 and Table 8. One can see that although the best results of different tasks (i.e., scar or edema segmentation) and metrics were achieved by different teams, UESTC and UBA performed consistently better than others or comparably with the best one or ground truth without statistical difference, except for the SPE of edema. By comparing these methods that achieved the best result in any metric or index, we conclude that the most recommendable strategies could be the two-stage coarse-to-fine procedure, model ensemble and data augmentation. Next, we will analyze both segmentation accuracy and clinical index results in detail, but we will mainly focus on the discussion on segmentation accuracy.

5.1.1. Result of segmentation accuracy

Table 7 presents the quantitative results of the evaluated algorithms for MyoPS. The average of Dice scores of the evaluated methods are 0.614 and 0.644 for scar and edema segmentation, respectively; and the average of ACC are 0.836 and 0.743 for scar and edema segmentation, respectively. Figure 4 provides the boxplots of Dice scores

Table 7. Summary of the quantitative evaluation results of scar and edema segmentation by the fifteen teams. Note that column Dice$^{\ominus}$reports the results excluding case #207, which contains no scar; and the average Dice changes from 0.614 ± 0.075 to 0.583 ± 0.072 if case #207 is included. Asterisk (*) indicates the method obtained statistically poorer results ($p < 0.01$) compared to the best performance in terms of corresponding metrics. ACC: accuracy; SEN: sensitivity; SPE: specificity.

Team	Scar				Edema			
	Dice$^{\ominus}$	ACC	SEN	SPE	Dice	ACC	SEN	SPE
UESTC	**0.708 ± 0.191**	0.870 ± 0.082	0.737 ± 0.185	0.925 ± 0.054	**0.731 ± 0.109**	**0.797 ± 0.095**	0.724 ± 0.134	$0.847 \pm 0.095^*$
UBA	0.701 ± 0.189	0.851 ± 0.075	**0.791 ± 0.175**	$0.867 \pm 0.070^*$	0.698 ± 0.129	0.762 ± 0.102	**0.748 ± 0.152**	$0.770 \pm 0.099^*$
NPU	0.681 ± 0.240	0.857 ± 0.105	0.734 ± 0.253	0.902 ± 0.096	0.709 ± 0.122	0.777 ± 0.112	0.703 ± 0.148	$0.819 \pm 0.133^*$
USTB	0.668 ± 0.255	0.852 ± 0.095	0.764 ± 0.257	$0.872 \pm 0.093^*$	0.688 ± 0.148	0.748 ± 0.135	0.741 ± 0.164	$0.736 \pm 0.184^*$
UHW	0.652 ± 0.195	0.848 ± 0.092	0.695 ± 0.232	$0.891 \pm 0.108^*$	$0.665 \pm 0.137^*$	$0.742 \pm 0.102^*$	0.722 ± 0.193	$0.744 \pm 0.169^*$
FZU	0.627 ± 0.215	$0.848 \pm 0.086^*$	0.632 ± 0.221	$0.931 \pm 0.043^*$	$0.686 \pm 0.123^*$	$0.777 \pm 0.084^*$	0.663 ± 0.151	$0.844 \pm 0.076^*$
NJUST	0.658 ± 0.241	**0.877 ± 0.074**	$0.642 \pm 0.269^*$	**0.952 ± 0.032**	$0.599 \pm 0.200^*$	0.771 ± 0.088	$0.501 \pm 0.211^*$	**0.943 ± 0.057**
CQUPT I	$0.637 \pm 0.227^*$	0.858 ± 0.084	$0.626 \pm 0.223^*$	0.938 ± 0.051	0.656 ± 0.138	0.766 ± 0.096	0.606 ± 0.179	$0.863 \pm 0.107^*$
LRDE	$0.617 \pm 0.233^*$	0.809 ± 0.142	$0.690 \pm 0.237^*$	$0.849 \pm 0.154^*$	$0.639 \pm 0.141^*$	$0.709 \pm 0.131^*$	0.698 ± 0.165	$0.716 \pm 0.199^*$
CQUPT II	$0.612 \pm 0.237^*$	$0.857 \pm 0.084^*$	$0.575 \pm 0.242^*$	0.951 ± 0.048	$0.725 \pm 0.110^*$	0.796 ± 0.100	$0.709 \pm 0.156^*$	$0.846 \pm 0.136^*$
HNU	$0.581 \pm 0.243^*$	$0.825 \pm 0.090^*$	$0.543 \pm 0.225^*$	0.923 ± 0.060	$0.619 \pm 0.166^*$	$0.751 \pm 0.110^*$	$0.544 \pm 0.190^*$	0.886 ± 0.081
Edin	$0.600 \pm 0.261^*$	$0.836 \pm 0.106^*$	$0.626 \pm 0.294^*$	0.925 ± 0.085	$0.603 \pm 0.182^*$	$0.733 \pm 0.113^*$	$0.572 \pm 0.220^*$	$0.843 \pm 0.127^*$
UBO	$0.595 \pm 0.244^*$	$0.806 \pm 0.096^*$	0.682 ± 0.281	$0.851 \pm 0.087^*$	$0.664 \pm 0.150^*$	$0.740 \pm 0.116^*$	0.716 ± 0.219	$0.760 \pm 0.142^*$
ITU	$0.595 \pm 0.229^*$	$0.824 \pm 0.098^*$	$0.632 \pm 0.258^*$	$0.898 \pm 0.079^*$	$0.612 \pm 0.160^*$	$0.739 \pm 0.112^*$	$0.575 \pm 0.199^*$	$0.838 \pm 0.102^*$
UOA	$0.493 \pm 0.251^*$	$0.817 \pm 0.110^*$	$0.453 \pm 0.273^*$	0.952 ± 0.036	$0.557 \pm 0.183^*$	$0.718 \pm 0.127^*$	$0.479 \pm 0.189^*$	0.881 ± 0.092
Average	0.614 ± 0.231	0.836 ± 0.096	0.643 ± 0.255	0.904 ± 0.088	0.644 ± 0.153	0.743 ± 0.112	0.645 ± 0.200	0.803 ± 0.148

Table 8. Summary of the clinical measures of automatic scar and edema segmentation from the fifteen teams and manual segmentation. Asterisk (*) indicates the method obtained statistically different results ($p < 0.01$) compared to ground truth. Here, the results with the largest coefficient of determination r^2 in the correlation study between the prediction and the ground truth are marked in bold.

Team	Scar			Edema		
	Transmurality	Surface area (cm^2)	Volume (cm^3)	Transmurality	Surface area (cm^2)	Volume (cm^3)
UESTC	0.727 ± 0.177	53.4 ± 23.7	1.32 ± 0.558	0.730 ± 0.175	150 ± 49.9	2.12 ± 0.697
UBA	0.601 ± 0.270	47.7 ± 32.8	1.14 ± 0.738	0.674 ± 0.196	146 ± 74.9	1.78 ± 0.836
NPU	$0.723 \pm 0.213^*$	58.3 ± 31.5	1.40 ± 0.659	0.719 ± 0.193	164 ± 73.0	2.15 ± 0.755
USTB	0.696 ± 0.240	61.5 ± 31.7	1.67 ± 0.878	0.720 ± 0.187	$169 \pm 72.3^*$	$2.73 \pm 1.32^*$
UHW	0.681 ± 0.240	57.1 ± 40.5	1.46 ± 0.985	0.738 ± 0.227	162 ± 89.1	2.52 ± 1.09
FZU	$0.721 \pm 0.234^*$	60.9 ± 33.0	1.68 ± 0.932	0.792 ± 0.202	$184 \pm 74.3^*$	$2.98 \pm 1.29^*$
NJUST	0.622 ± 0.257	43.4 ± 28.6	1.11 ± 0.618	**$0.827 \pm 0.147^*$**	136 ± 57.1	**2.28 ± 0.815**
CQUPT I	**0.664 ± 0.245**	42.5 ± 25.2	$1.02 \pm 0.575^*$	$0.826 \pm 0.148^*$	**139 ± 65.8**	2.12 ± 0.926
LRDE	$0.763 \pm 0.232^*$	62.1 ± 37.0	1.59 ± 0.818	$0.849 \pm 0.142^*$	170 ± 81.3	2.58 ± 1.11
CQUPT II	0.691 ± 0.221	51.8 ± 33.2	1.11 ± 0.601	0.675 ± 0.200	157 ± 80.0	1.80 ± 0.832
HNU	0.673 ± 0.178	42.9 ± 18.9	1.08 ± 0.388	0.583 ± 0.301	131 ± 47.2	$1.54 \pm 0.544^*$
Edin	0.707 ± 0.169	**59.4 ± 30.3**	**$1.73 \pm 0.854^*$**	0.786 ± 0.178	162 ± 53.6	2.47 ± 0.820
UBO	$0.762 \pm 0.210^*$	62.7 ± 32.1	1.56 ± 0.703	0.598 ± 0.308	174 ± 70.1	2.35 ± 0.828
ITU	$0.749 \pm 0.186^*$	54.1 ± 27.4	1.30 ± 0.680	0.714 ± 0.202	152 ± 68.7	1.82 ± 0.797
UOA	0.645 ± 0.216	$32.0 \pm 20.2^*$	$0.750 \pm 0.415^*$	0.716 ± 0.202	117 ± 55.8	$1.37 \pm 0.550^*$
Ground truth	0.606 ± 0.224	53.6 ± 30.7	1.44 ± 0.893	0.658 ± 0.203	149 ± 64.9	2.24 ± 0.874

and ACC of each team. The best Dice scores (0.708 ± 0.191 and 0.731 ± 0.109 for scar and edema segmentation, respectively) were both achieved by UESTC; but the best ACC, SEN and SPE were accomplished by NJUST & UESTC, UBA, and NJUST, respectively. In general, the evaluated methods achieved worse performance for scar segmentation than for edema segmentation in terms of Dice, but not in terms of the other three metrics. In fact, the same metric and value could often refer to different degrees of clinical acceptability for different tasks, depending on the size and shape of the target object and the complexity of form (Li et al., 2020a). For example, Dice tends to be more sensitive to the small deviations in segmentation for small sparse objects than for large, compact objects, which may explain the better Dice results for edema, which has a larger volume and is less patchy from T2 CMR images. This conclusion is more evident when we compare the inter-observer of scar and edema in terms of Dice (0.569 ± 0.198 vs. 0.701 ± 0.168). Interestingly, one algorithm could perform well in terms of Dice in one pathology but not necessarily in another, for example, CQUPT II excelled in edema segmentation but performed poorly in scar segmentation. Moreover, these algorithms generally showed different segmentation capabilities across different slices. Figure 5 illustrates the performances of each team on basal, middle and apical slices, respectively. One can see that the pathologies in the middle areas were much easier to segment than that at either end slice, while the results in the basal areas usually deviated significantly. In general, the algorithms that performed well overall also presented a similar competitive performance on challenging apical and/ or basal slices (please refer to Section 5.2 for further analysis). For more details on the performance of each team, please refer to the Supplementary Material, where we further provided the HD and precision values.

Figure 6 provides boxplots of Dice and ACC from the

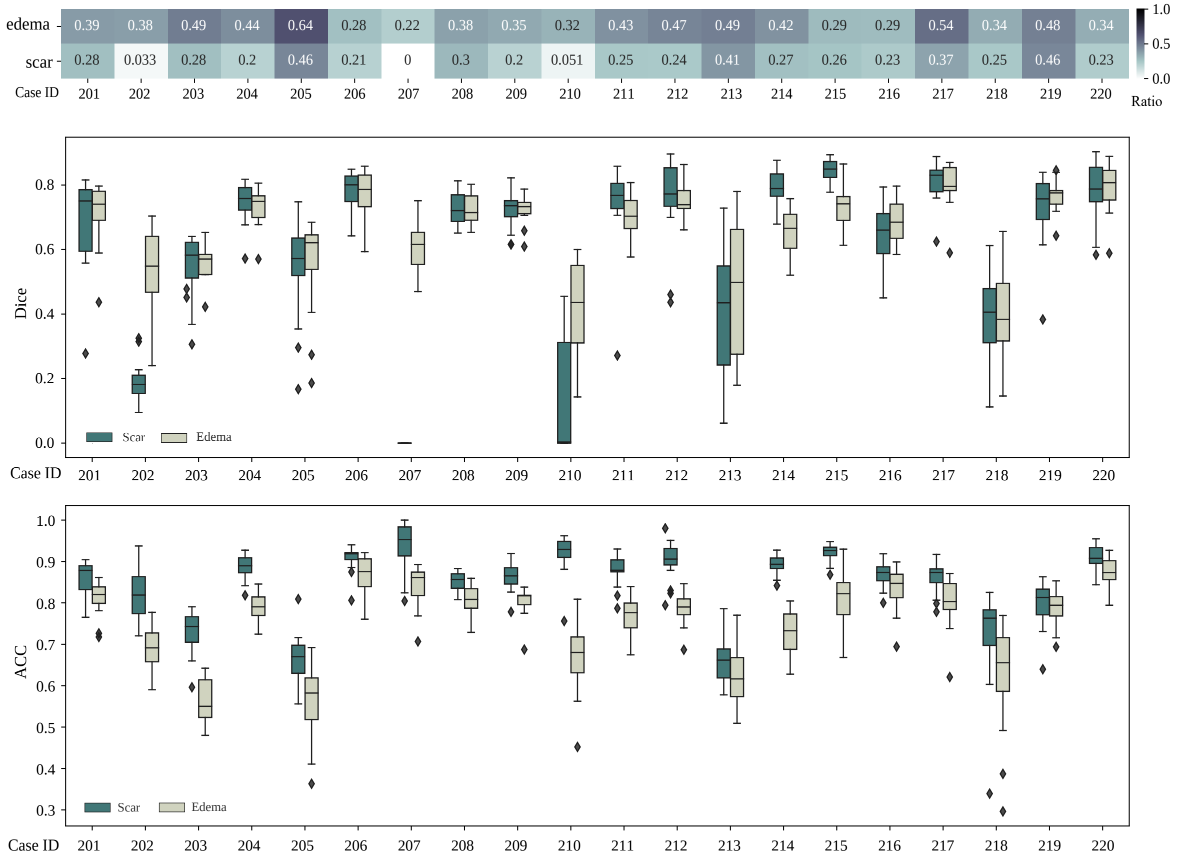


Figure 6. Ratio of pathologies to myocardium and the average performance, *i.e.*, Dice and ACC, of the evaluated algorithms on each test case.

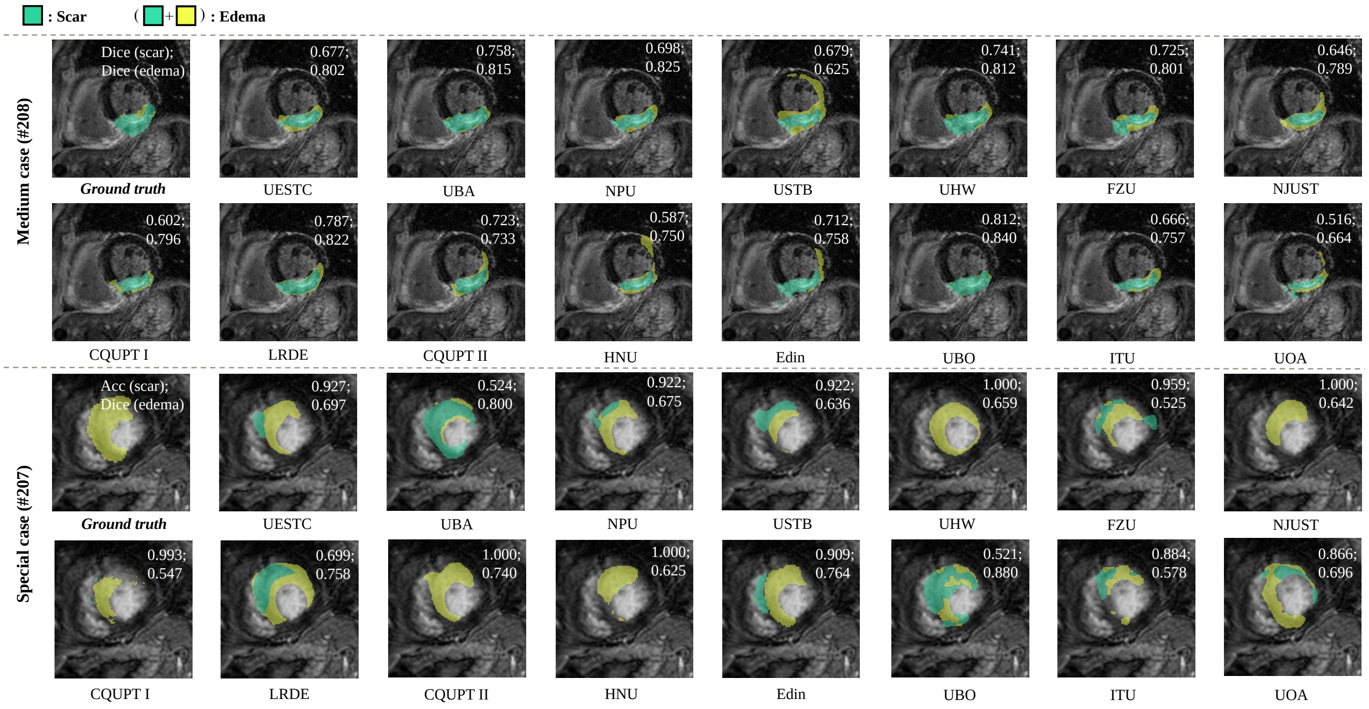


Figure 7. Visualization of the segmentation results of the median and special case by the evaluated methods from each team. The median cases were from the test set in terms of mean Dice of scar and edema by the fifteen methods, while the special case contain no scar. Note that the segmentation masks are superimposed on the LGE CMR image which is used for anatomy reference.

evaluated algorithms on each test case, where the ratios of pathologies to myocardium are provided for reference. One can see that there exist large variations of the performance among different cases in terms of both Dice and ACC. The position, shape and extent of pathologies all could affect the performance, and will be analyzed in Sections 5.2, 5.3 and 5.4, respectively. Particularly, the slices without pathology could confuse the algorithms, of which most segment pathology slice-wisely, *i.e.,* an algorithm segments the CMR images slice by slice instead of as a whole volume. Note that the situation of a slice without pathology is more often than that of a subject without pathology. In the test data, only one special subject (#207) has no scar (see Figure 7). For this case, it could easily induce a Dice score of zero for the evaluation of an algorithm if it misclassifies even only one voxel, according to its definition.

Figure 7 visualizes the segmentation results of the middle slice of the special case (#207) and the median case (#208). Most methods achieved good results for the median case (#208), although some contained patchy noises. Specifically, the results of median case by USTB, HNU, Edin and UOA contain significant amount of outliers of edema, and parts of scars are evidently mis-classified into edema by USTB, FZU and NJUST. For the special cases (#207), false positives of scar classification were the major errors. Only UHW, NJUSTM CQUPT II and HUN contained no false positive of scar classification, and thus were evaluated with ACC of 1.000; UBA and UBO mistook edema as scars, but still obtained high Dice scores for the segmentation of edema which includes both scarring and peri-infarct region. Nevertheless, this indicates the difficulty of differentiating the scars and peri-infarct regions, which is currently out of the scope of this study.

5.1.2. Result of clinical indices

Table 8 compares the average clinical measure results of the evaluated algorithms and manual segmentation. One can see that most teams could obtain similar values compared to ground truth, which can be further proved in the correlation studies. Readers can refer to Figures 1, 2 and 3 in the Supplementary Material for more details on the correlation analysis between the clinical indices obtained by each method and the ground truth. We found that in general the top-performed teams in terms of Dice presented good consistency with the ground truth on the clinical index measurement. This is reasonable as the measurement was calculated based on the segmentation results. Nevertheless, there still exists inconsistency between the segmentation accuracy and surface area/ volume. For example, the edema surface area and volume of USTB were significantly different from the ground truth, even though its segmentation accuracy was promising. This could be due to the fact that the slice spacing of images is quite large, which may introduce non-negligible evaluation biases. Hence, even though several methods obtained inaccurate segmentation, such as UBO and ITU, their area or volume predictions could still present a small difference from the ground truth.

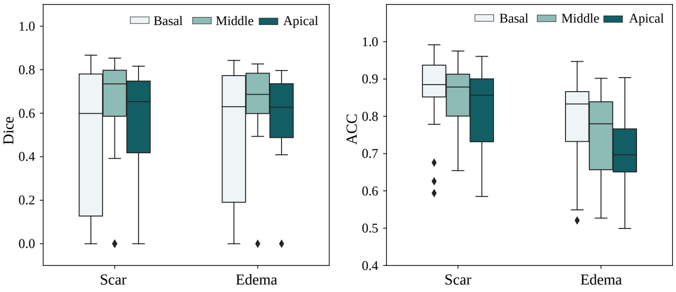


Figure 8. The boxplots of the average Dice and ACC of pathology segmentation with respect to different slice positions.

Moreover, most methods can obtain relatively accurate surface area and volume, but performed badly on the transmurality prediction, as proved in the correlation studies. For example, the coefficient of determination r^2 is quite low for all methods in terms of “transmurality of edema” (r^2<0.29). Compared to surface area and volume, transmurality is a more local measurement, as it considers the location of pathology. One can see that the prediction of transmurality is more consistent with the segmentation accuracy than surface area and volume across all methods. For example, the top-performed methods in terms of Dice generally have slopes closer to 1 in the correlation analysis of transmurality. In contrast, for surface area and volume, both slopes and r^2 show less variation among these teams, which complicates the clinical translatability evaluation.

5.2. Performance versus position of pathology

To analyze the correlation between the performance of MyoPS and the position of pathologies, we generated the boxplots of Dice and ACC of MyoPS at different slice positions, as shown in Figure 8. It is evident that the results of different slices were different, representing varying types of challenges, which is consist with aforementioned observation. From the Dice results, which represent overlap of pathologies from two segmentation results, the best performance was observed in the middle slices. This is reasonable as the ventricles in apical and basal slices usually exhibit more irregular and small-shaped pathologies, which may introduce additional challenges for the segmentation. Moreover, the performances of basal present particularly poor results with low mean values and large variance. This could be due to less presence of pathologies in basal slices, which is visualized in Figure 9, and a few poor cases inducing particularly low segmentation Dice. From the ACC results, whose calculation considers the classification on both of the positives and negatives of pathological segmentation, Basal has higher-valued box plots. This could be again attributed to the rare cases of pathologies occurred in basal slices, which should be discussed below.

Figure 9 visualizes the distribution maps of pathologies from the 20 test subjects, and the SEN and SPE maps of

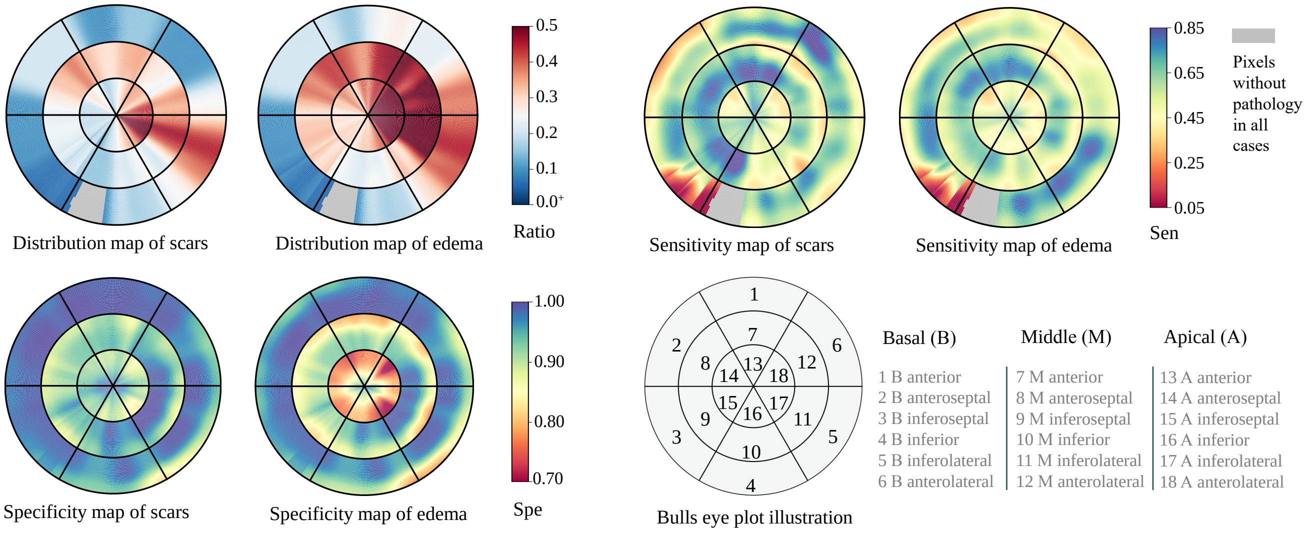


Figure 9. Bull's eye plots of pathology maps and mean segmentation performance with respect to different segments. Note that the grey-colored regions in distribution maps and sensitivity maps indicate none of the cases has pathology in the positions. The color map scales of SEN and SPE are different, and readers should take the difference into consideration when interpreting them.

MyoPS using 2D bull's eye plots. As there are various slice numbers among different cases, we normalized the slice positions for each case referring to Liu et al. (2016) and Zhuang et al. (2011), and the maps were averaged from the set of segments from different slices, subjects and different classification results by the benchmarked methods (only for SPE and SEN maps). From the distribution maps, one can see that scars mainly occur in the inferolateral regions and anterior segments of middle slices in this test dataset; and edema extends to more regions of basal slices and almost all segments of apical and middle slices. From the SEN (true positive rate) and SPE (true negative rate) maps of pathologies, one can observe that the regional values of SPE were generally higher than that of SEN, which could be due to the definition of classification, namely the pixels not segmented as pathologies by an algorithm were regarded as negatives by default.

For scar segmentation, we found that the SEN values were higher in middle septum segments. It could be attributed to the good contrast from C0 and T2 for myocardium segmentation of septum, leading to an easier scar segmentation task from LGE myocardium. By contrast, the low values in SEN maps of both scar and edema are distributed in the area of basal inferoseptal segments, where there should be few cases having pathology, and the models to segment these areas were under trained. This explains the particularly low Dice of the first quartile in the Dice boxplot of Basal slices in Figure 8. Similarly, one can observe from both of the sensitivity and specificity maps that the performance of MyoPS on near-endocardium areas was generally better than that on the near-epicardium regions. Here, near-endocardium and near-epicardium areas refer to the regions near the inner and outer layer of the heart, respectively (inner and outer race of bull's eye plots in Figure 9) (Virmani et al., 1990). This could be due to the better contrast in the areas between myocardium and ventricular blood pools than that between myocardium and adjacent tissues (liver and lung) in all the three CMR sequences.

5.3. *Performance versus shape of pathology*

Figure 10 presents the correlations between the mean segmentation accuracy (Dice or ACC) and the shape of pathologies. Here, we employed compactness to quantify the shape of pathologies in a slice, which is defined to the ratio of the area of an object to the area of a circle with the same perimeter (Bogaert et al., 2000). As a circle is regarded as the object with the most compact shape, the measure normally takes a maximum value of 1 for a circle. One can see that there are positive correlations between the pathology shape and the performance, which is evident for ACC though marginal for Dice. This could reveal that the pathologies with asymmetric shapes could be more easily mis-classified by the benchmarked algorithms.

5.4. *Performance versus extent of pathology*

Figure 11 presents the correlations between the mean segmentation accuracy (Dice or ACC) and the extent of pathologies. We analyzed the correlations from two perspectives, i.e., subject-wise correlation and slice-wise correlation, respectively. The subject-wise computation refers to the computation on a subject, which may include 3-5 slices, while the slice-wise computation refers to the computation on a slice. Compared to subject-wise computation, slice-wise computation may be more robust as more instances can be included. Also, the distribution of pathology is heterogeneous across slices (see Figure 9). One can see that the ACC values of the pathology segmentation were negatively correlated with extent of pathologies in

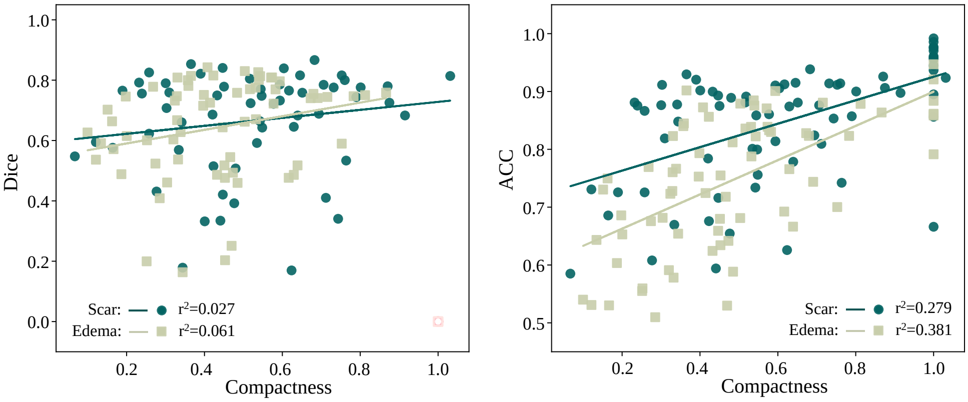


Figure 10. The scatter point plots and correlation between the performance of pathology segmentation with respect to the compactness of the pathologies. Note that the slices without pathologies, indicated by light pink hollow scatter points, are excluded in the computation of correlation with Dice.

both the subject-wise and slice-wise studies, but no evident correlation in terms of Dice was observed in either study. Note that non-pathological myocardium pixels were defined as negatives by default, and the cases having large area of negatives tended to have higher ACC values due to this definition. In contrast, the variation of pathology sizes did not have an evident influence on the final performance in terms of Dice. Overall, the subject-wise and slice-wise correlations were consistent, especially in terms of ACC.

6. Discussion

6.1. Variation of manual segmentation versus performance and variation of automatic segmentation

All the reviewed algorithms were based on supervised learning, so their performance could depend on the quality of labels. For MyoPS, the inter-observer variability is generally large due to the poor image quality and small volume of targets. In other words, different experts could offer variable manual segmentation results under the influence of background knowledge and levels of expertise of raters. To analyze the effect of inter-observer variations on the segmentation performance of automated algorithms, we first performed a correlation analysis between the inter-observer variations and the average performance of all submitted models from participants; and we further analyzed the relationship between the inter-observer variations and the inter-participant variations. Here, inter-observer/ participant variations are defined to the average Dice$^{\circleddash}$ or Dice scores between different segmentation results.

Figure 12 presents results of correlation studies. The inter-observer variation can be considered as a representation of uncertainty of manual segmentation, which may reveal the difficulties of segmentation. However, the average Dice scores of the automatic models were not strongly relevant to inter-observer variations. Note that the high r^2 value for Scar could be attributed to the three special cases highlighted by the red arrows in Figure 12. Similarly, the inter-participant variations can be regarded as the uncertainties of automatic models, which nevertheless had weak correlation to the uncertainties of manual segmentation in this study.

6.2. Discussion of pre-alignment and MyoPS of CMR

We visually checked each case the alignment result of the three-sequence CMR, and assigned a score, ranging from 0 to 5, to represent the quality of alignment. The score of 5 indicates perfect alignment, 1 to 4 denotes misalignment from severe to marginal, and 0 suggests completely failed alignment. The majority of cases were well aligned, as Figure 13 presents, and the average scores were 4.50 ± 0.931 and 4.68 ± 0.789 in the training set and test set, respectively.

One can see that the alignment score is ordinal data in a non-Gaussian distribution. Therefore, to analyze the effect of pre-alignment on the automatic segmentation, we performed a Spearman's rank correlation analysis (Sedgwick, 2014), on the alignment scores and average Dice$^{\circleddash}$(for scar) or Dice (for edema). The Spearman coefficient for scar and edema segmentation were respectively -0.207 (p-value = 0.740) and -0.143 (p-value = 0.252), indicating no evidence of significant relationship between these figures.

The limitation comes from the fact that majority of the cases were well pre-aligned, followed by the segmentation combining the three-sequence MRI. Hence, future studies should include the original images without alignments for both training and testing of DL-based models. Also, since DL-based method has a great potential to achieve combined computing of simultaneous registration and segmentation. Such strategy of combined computing for MyoPS could be further explored in the future.

6.3. Discussion of evaluation metrics and ranking

As Table 7 presents, different evaluation metrics could lead to different ranking results for an algorithm or team, indicating potential limitation of the metrics and unfairness of ranking. Particularly, classification metrics could be misleading for assessment of semantic segmentation. For example, ACC is sensitive to the volume of targets,

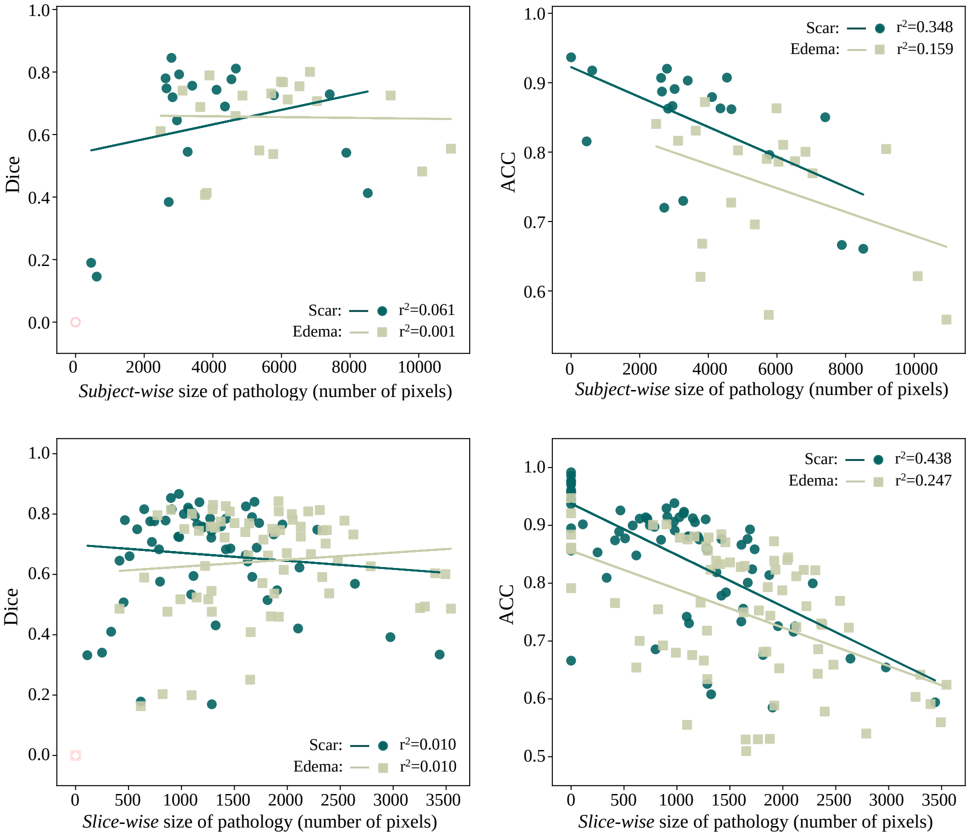


Figure 11. The scatter point plots and correlation between the performance of pathology segmentation with respect to the size of the pathologies for patient-wise and slice-wise computation. Again, note that the subjects without pathology, indicated by light pink hollow scatter points, are excluded in the computation of correlation with Dice.

and Dice score will fail to act as a metric when the target manual segmentation is none, as discussed in Section 3.2 and Figure 7. Moreover, the classification metrics may also be misleading when considering its accuracy of clinical index calculation. Therefore, we argued that methodology survey and case studies could be more valuable and convincing for benchmark, than ranking the methods according to the figures of evaluation.

6.4. Clinical translatability

There are several key points to consider in determining how current MyoPS algorithms fit into the clinical workflow. (1) Accuracy. Several clinical indices, such as scar/ edema area, transmurality and positions, are very important for the diagnosis and surgical assessment of MI. To obtain this information, accurate segmentation of scar and edema is required. (2) Generalizability. Due to different data acquisition protocols, test data may differ significantly in quality or appearance from training data, which can lead to significant degradation in performance. Therefore, the trained model requires sufficient generalization to apply to test data outside the training data distribution. (3) Time efficiency. Tolerable computation time is quite important in a real-time system.

The best performances on scar and edema were respectively 0.708±0.191 and 0.731±0.109 in terms of Dice score. The results have already presented promising potential, particularly when comparing with the inter-observer Dice scores of scars and edema, which are 0.569 ± 0.198 and 0.701 ± 0.168, respectively. We also calculated the clinical indices for the evaluated algorithms, most of which were comparable to the ground truth results. As no extra dataset from other center/ vendor was provided for evaluation, we can not analyze the generalizability of submitted algorithms. The processing in test stage consists of data preprocessing, prediction, and post-processing, which all contribute to the spent time. In this challenge, for data preprocessing, we have registered the three sequences in advance, which makes the subsequent segmentation task easier. However, registration is indeed time-consuming (7.57±2.51 min on a system with a i7-7700k 4.2GHZ CPU and 16GB RAM). Therefore, how to effectively fuse the information in these sequences needs to be investigated in the future.

6.5. Limitation and future prospects

There is a gap between technique design and clinical prior knowledge of manual segmentation. In the clinic, the criteria used to determine the presence of scar/ edema partially relies on the anatomical knowledge of the myocardium. Specifically, the LV myocardium can be divided into unified 18 segments, displayed on a circumfer-

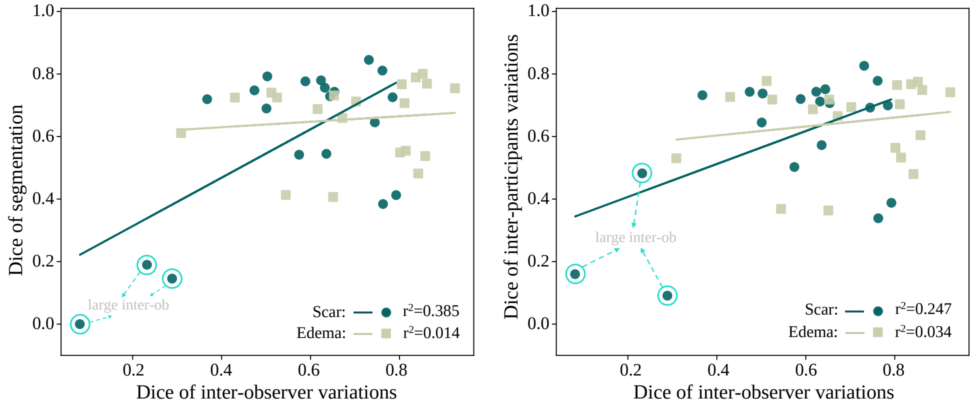


Figure 12. The scatter point plots and correlations between the inter-observer variations and the average performance in terms of Dice, and the inter-participant variations. Here, the dash arrows identify the cases with large inter-observer variations (low Dice values).

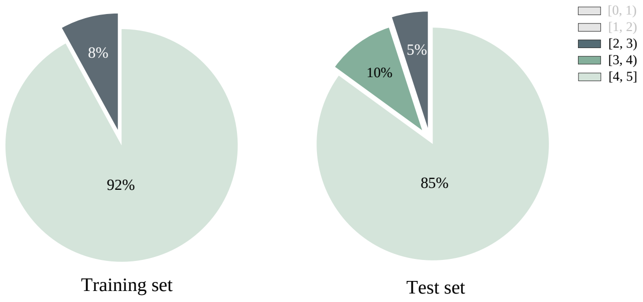


Figure 13. The distribution of the pre-alignment scores (scored from 0 to 5) in training and test sets of MyoPS dataset.

ential polar plot, as shown in Figure 9. Despite variability in the coronary artery blood supply to myocardium, it was believed that every segment can be supplied by specific coronary artery territories (Cerqueira et al., 2002). There are three major coronary arteries, each of which supplies its own specific coronary artery territories. For example, segments 1, 2, 7, 8, 13 and 14 (left area in bullseyes plots) are supplied by the left anterior descending coronary artery. With these artery territory knowledge, it is known that the ischemia area (including the scarring/ edema area) generally does not cross two territories, since the successive ischemia is commonly caused by a single vascular occlusion. Hence, the predicted area across territory should be penalized. However, current methods did not consider this anatomical knowledge of pathology when designing their algorithms. Therefore, in the future we expect more research on novel methodologies to combine this anatomical knowledge into their framework for more accurate and clinical-related MyoPS results. Moreover, in this challenge only the short-axis image is used for analysis, while the complementary information from long axis is also crucial in clinical practice for scar localization (Chan et al., 2006). In the future, we expect the methods to combine multi-view CMR images for this task.

7. Conclusion

This paper surveys the submitted works from the MyoPS challenge, which provides 45 sets of three-sequence CMR images. Fifteen algorithms were benchmarked for comparisons, and their methodologies and segmentation performance were then analyzed and examined. To the best of our knowledge, this is the first work to evaluate simultaneous scar and edema segmentation combining multi-source images. All the benchmarked methods fully utilized the complementary information of the pre-aligned three-sequence CMR images. However, none of them considered the problem of misalignment between the images or used other data sources. These are the main limitations of this study, and the problems remain to be further explored. In the future, we expect more research on simultaneous registration and fusion of multi-source data for pathology segmentation. Note that the data and evaluation tool continue as ongoing resources for the community.

Author contributions

XZ initialized the challenge and provided all the resources; XZ, LL, FW, SW and XL collected the materials and composed the manuscript. CM, JZ, YL, ZZ, SZ, MA, HJ, XZ, LW, TA, EA, ZZ, FL, JM, XY, EP, IO, SB, WL, KP, ST, LS, GW, MY, GL, YX, SE were participants of the MyoPS challenge. The participants described their algorithms and segmentation results for evaluation, and contributed equally to this paper. All the authors have read and approved the publication of this work.

Acknowledgement

This work was funded by the National Natural Science Foundation of China (grant no. 62111530195 and 61971142).

References

Ankenbrand, M.J., Lohr, D., Schreiber, L.M., 2020. Exploring ensemble applications for multi-sequence myocardial pathology segmentation, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 60–67.

Arega, T.W., Bricq, S., 2020. Automatic myocardial scar segmentation from multi-sequence cardiac MRI using fully convolutional densenet with inception and squeeze-excitation module, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 102–117.

Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S., 2019. Bi-directional convlstm u-net with densley connected convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 .

Baron, N., Kachenoura, N., Beygui, F., Cluze, P., Grenier, P., Herment, A., Frouin, F., 2008. Quantification of myocardial edema and necrosis during acute myocardial infarction, in: 2008 Computers in Cardiology, IEEE. pp. 781–784.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Transactions on Medical Imaging 37, 2514–2525.

Bogaert, J., Rousseau, R., Van Hecke, P., Impens, I., 2000. Alternative area-perimeter ratios for measurement of 2d shape compactness of habitats. Applied Mathematics and Computation 111, 71–85.

Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Transactions on Medical Imaging .

Cerqueira, M.D., Weissman, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennell, D.J., Rumberger, J.A., Ryan, T., et al., 2002. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. Circulation 105, 539–542.

Chan, J., Hanekom, L., Wong, C., Leano, R., Cho, G.Y., Marwick, T.H., 2006. Differentiation of subendocardial and transmural infarction using two-dimensional strain rate imaging to assess short-axis and long-axis myocardial function. Journal of the American College of Cardiology 48, 2026–2033.

Delgado, V., Van Bommel, R.J., Bertini, M., Borleffs, C.J.W., Marsan, N.A., Ng, A.C., Nucifora, G., Van De Veire, N.R., Ypenburg, C., Boersma, E., et al., 2011. Relative merits of left ventricular dyssynchrony, left ventricular lead position, and myocardial scar to predict long-term survival of ischemic heart failure patients undergoing cardiac resynchronization therapy. Circulation 123, 70–78.

Du, J., Li, W., Lu, K., Xiao, B., 2016. An overview of multi-modal medical image fusion. Neurocomputing 215, 3–20.

Elif, A., Ilkay, O., 2020. Accurate myocardial pathology segmentation with residual u-net, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 128–137.

Friedrich, M.G., 2010. Myocardial edema—a new clinical entity? Nature Reviews Cardiology 7, 292–296.

Gao, H., Kadir, K., Payne, A.R., Soraghan, J., Berry, C., 2013. Highly automatic quantification of myocardial oedema in patients with acute myocardial infarction using bright blood T2-weighted CMR. Journal of Cardiovascular Magnetic Resonance 15, 1–12.

Jiang, H., Wang, C., Chartsias, A., Tsaftaris, S.A., 2020. Max-fusion u-net for multi-modal pathology segmentation with attention and dynamic resampling, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 68–81.

Kadir, K., Gao, H., Payne, A., Soraghan, J., Berry, C., 2011. Automatic quantification and 3d visualisation of edema in cardiac MRI, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. pp. 8021–8024.

Karim, R., Bhagirath, P., Claus, P., Housden, R.J., Chen, Z., Karimaghaloo, Z., Sohn, H.M., Rodríguez, L.L., Vera, S., Albà, X., et al., 2016. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement MR images. Medical Image Analysis 30, 95–107.

Karim, R., Blake, L.E., Inoue, J., Tao, Q., Jia, S., Housden, R.J., Bhagirath, P., Duval, J.L., Varela, M., Behar, J., et al., 2018. Algorithms for left atrial wall segmentation and thickness–evaluation on an open-source CT and MRI image database. Medical Image Analysis 50, 36–53.

Karim, R., Housden, R.J., Balasubramaniam, M., Chen, Z., Perry, D., Uddin, A., Al-Beyatti, Y., Palkhi, E., Acheampong, P., Obom, S., et al., 2013. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. Journal of Cardiovascular Magnetic Resonance 15, 105.

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al., 2021. Chaos challenge-combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis 69, 101950.

Kidambi, A., Mather, A.N., Swoboda, P., Motwani, M., Fairbairn, T.A., Greenwood, J.P., Plein, S., 2013. Relationship between myocardial edema and regional myocardial function after reperfused acute myocardial infarction: an MR imaging study. Radiology 267, 701–708.

Kurzendorfer, T., Breininger, K., Steidl, S., Brost, A., Forman, C., Maier, A., 2018. Myocardial scar segmentation in LGE-MRI using fractal analysis and random forest classification, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE. pp. 3168–3173.

Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., de la Rosa, E., Cochet, A., Cottin, Y., Ginhac, D., et al., 2020. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI. Data 5, 89.

Li, F., Li, W., 2020. Dual-path feature aggregation network combined multi-layer fusion for myocardial pathology segmentation with multi-sequence cardiac MR, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 146–158.

Li, J., Udupa, J.K., Tong, Y., Wang, L., Torigian, D.A., 2020a. Linsem: Linearizing segmentation evaluation metrics for medical images. Medical Image Analysis 60, 101601.

Li, L., Ding, W., Huang, L., Zhuang, X., Grau, V., 2022a. Multi-modality cardiac image computing: A survey. arXiv preprint arXiv:2208.12881 .

Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X., 2022b. Medical image analysis on left atrial LGE MRI for atrial fibrillation studies: A review. Medical Image Analysis 77, 102360.

Li, W., Wang, L., Qin, S., 2020b. CMS-UNet: Cardiac multi-task segmentation in MRI with a u-shaped network, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 92–101.

Liu, D., Hu, K., Nordbeck, P., Ertl, G., Störk, S., Weidemann, F., 2016. Longitudinal strain bull's eye plot patterns in patients with cardiomyopathy and concentric left ventricular hypertrophy. European Journal of Medical Research 21, 1–12.

Liu, Y., Zhang, M., Zhan, Q., Gu, D., Liu, G., 2020. Two-stage method for segmentation of the myocardial scars and edema on multi-sequence cardiac magnetic resonance, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 26–36.

Lu, Y., Yang, Y., Connelly, K.A., Wright, G.A., Radau, P.E., 2012. Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images. Quantitative imaging in medicine and surgery 2, 81.

Ma, J., 2020. Cascaded framework with complementary CMR information for myocardial pathology segmentation, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 159–166.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. ISLES 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269.

Martín-Isla, C., Asadi-Aghbolaghi, M., Gkontra, P., Campello, V.M., Escalera, S., Lekadir, K., 2020. Stacked BCDU-Net with semantic CMR synthesis: Application to myocardial pathology segmentation challenge, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 1–16.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 1993–2024.

Moccia, S., Banali, R., Martini, C., Muscogiuri, G., Pontone, G., Pepi, M., Caiani, E.G., 2019. Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images. Magnetic Resonance Materials in Physics, Biology and Medicine 32, 187–195.

Moghari, M.H., Pace, D.F., Akhondi-Asl, A., Powell, A.J., 2016. HVSMR 2016: MICCAI workshop on whole-heart and great vessel segmentation from 3D cardiovascular MRI in congenital heart disease. `http://segchd.csail.mit.edu/index.html`.

Ørn, S., Manhenke, C., Anand, I.S., Squire, I., Nagel, E., Edvardsen, T., Dickstein, K., 2007. Effect of left ventricular scar size, location, and transmurality on left ventricular remodeling with healed myocardial infarction. The American journal of cardiology 99, 1109–1114.

Ortiz-Pérez, J.T., Meyers, S.N., Lee, D.C., Kansal, P., Klocke, F.J., Holly, T.A., Davidson, C.J., Bonow, R.O., Wu, E., 2007. Angiographic estimates of myocardium at risk during acute myocardial infarction: validation study using cardiac magnetic resonance imaging. European heart journal 28, 1750–1758.

Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2337–2346.

Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C., et al., 2015. Right ventricle segmentation from cardiac MRI: a collation study. Medical Image Analysis 19, 187–202.

Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K., 1987. Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing 39, 355–368.

Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G., 2009. Evaluation framework for algorithms segmenting short axis cardiac MRI. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge 49.

Raman, S.V., Simonetti, O.P., Winner, M.W., Dickerson, J.A., He, X., Mazzaferri, E.L., Ambrosio, G., 2010. Cardiac magnetic resonance with edema imaging identifies myocardium at risk and predicts worse outcome in patients with non–ST-segment elevation acute coronary syndrome. Journal of the American College of Cardiology 55, 2480–2488.

Rohlfing, T., Maurer, C.R., 2006. Shape-based averaging. IEEE Transactions on Image Processing 16, 153–161.

Ruder, T.D., Ebert, L.C., Khattab, A.A., Rieben, R., Thali, M.J., Kamat, P., 2013. Edema is a sign of early acute myocardial infarction on post-mortem magnetic resonance imaging. Forensic Science, Medicine, and Pathology 9, 501–505.

Sandfort, V., Kwan, A.C., Elumogo, C., Vigneault, D.M., Symons, R., Pourmorteza, A., Rice, K., Davies-Venn, C., Ahlman, M.A., Liu, C.Y., et al., 2017. Automatic high-resolution infarct detection using volumetric multiphase dual-energy CT. Journal of cardiovascular computed tomography 11, 288–294.

Schuijf, J.D., Kaandorp, T.A., Lamb, H.J., van der Geest, R.J., Viergever, E.P., van der Wall, E.E., de Roos, A., Bax, J.J., 2004. Quantification of myocardial infarct size and transmurality by contrast-enhanced magnetic resonance imaging in men. The American journal of cardiology 94, 284–288.

Sedgwick, P., 2014. Spearman's rank correlation coefficient. Bmj 349.

Suinesiaputra, A., Cowan, B.R., Finn, J.P., Fonseca, C.G., Kadish, A.H., Lee, D.C., Medrano-Gracia, P., Warfield, S.K., Tao, W., Young, A.A., 2011. Left ventricular segmentation challenge from cardiac MRI: a collation study, in: International Workshop on Statistical Atlases and Computational Models of the Heart, pp. 88–97.

Takahashi, R., Matsubara, T., Uehara, K., 2019. Data augmentation using random image cropping and patching for deep cnns. IEEE Transactions on Circuits and Systems for Video Technology 30, 2917–2931.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR. pp. 6105–6114.

Tao, Q., Milles, J., Zeppenfeld, K., Lamb, H.J., Bax, J.J., Reiber, J.H., van der Geest, R.J., 2010. Automated segmentation of myocardial scar in late enhancement MRI using combined intensity and spatial information. Magnetic Resonance in Medicine 64, 586–594.

Tao, Q., Piers, S.R., Lamb, H.J., Zeppenfeld, K., van der Geest, R.J., 2015. Myocardial scar surface area identified by LGE MRI is an independent predictor of mortality in post-infarction patients. Journal of Cardiovascular Magnetic Resonance 17, 1–2.

Thygesen, K., Alpert, J.S., White, H.D., 2008. Universal definition of myocardial infarction. European Heart Journal 29, 1209.

Tobon-Gomez, C., Geers, A.J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z., et al., 2015. Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets. IEEE Transactions on Medical Imaging 34, 1460–1473.

Vall, J.M., Lemaitre, G., 2016. I2cvb: initiative for collaborative computer vision benchmark. `https://i2cvb.github.io/`.

Virmani, R., Forman, M., Kolodgie, F., 1990. Myocardial reperfusion injury. histopathological effects of perfluorochemical. Circulation 81, IV57–68.

Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al., 2020. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical Image Analysis 67, 101832.

Xu, C., Xu, L., Gao, Z., Zhao, S., Zhang, H., Zhang, Y., Du, X., Zhao, S., Ghista, D., Liu, H., et al., 2018. Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture. Medical Image Analysis 50, 82–94.

Yu, H., Zha, S., Huangfu, Y., Chen, C., Ding, M., Li, J., 2020. Dual attention u-net for multi-sequence cardiac MR images segmentation, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 118–127.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31, 1116–1128.

Zabihollahy, F., White, J.A., Ukwatta, E., 2019. Convolutional neural network-based approach for segmentation of left ventricle myocardial scar from 3d late gadolinium enhancement MR images. Medical physics 46, 1740–1751.

Zhai, S., Gu, R., Lei, W., Wang, G., 2020. Myocardial edema and scar segmentation using a coarse-to-fine framework with weighted ensemble, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 49–59.

Zhang, J., Xie, Y., Liao, Z., Verjans, J., Xia, Y., 2020a. Efficientseg: A simple but efficient solution to myocardial pathology segmentation challenge, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 17–25.

Zhang, X., Noga, M., Punithakumar, K., 2020b. Fully automated deep learning based segmentation of normal, infarcted and edema regions from multiple cardiac MRI sequences, in: Myocardial

Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 82–91.

Zhang, Z., Liu, C., Ding, W., Wang, S., Pei, C., Yang, M., Huang, L., 2020c. Multi-modality pathology segmentation framework: Application to cardiac magnetic resonance images, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 37–48.

Zhao, Z., Boutry, N., Puybareau, É., 2020. Stacked and parallel U-nets with multi-output for myocardial pathology segmentation, in: Myocardial Pathology Segmentation Combining Multi-Sequence CMR Challenge, Springer. pp. 138–145.

Zhuang, X., 2019. Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2933 – 2946.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. Medical Image Analysis 58, 101537.

Zhuang, X., Shi, W., Duckett, S., Wang, H., Razavi, R., Hawkes, D., Rueckert, D., Ourselin, S., 2011. A framework combining multi-sequence MRI for fully automated quantitative analysis of cardiac global and regional functions, in: International Conference on Functional Imaging and Modeling of the Heart, Springer. pp. 367–374.

Zhuang, X., Xu, J., Luo, X., Chen, C., Ouyang, C., Rueckert, D., Campello, V.M., Lekadir, K., Vesal, S., RaviKumar, N., et al., 2022. Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge. Medical Image Analysis 81, 102528.

Supplementary Material: Benchmark of Myocardial Pathology Segmentation Combining Three-Sequence Cardiac Magnetic Resonance Images

Lei Li[1,2†], Fuping Wu[1†], Sihan Wang[1†], Xinzhe Luo[1], Carlos Martín-Isla[3], Shuwei Zhai[5], Jianpeng Zhang[6], Yanfei Liu[7], Zhen Zhang[9], Markus J. Ankenbrand[10], Haochuan Jiang[11,12], Xiaoran Zhang[13], Linhong Wang[15], Tewodros Weldebirhan Arega[16], Elif Altunok[17], Zhou Zhao[18], Feiyan Li[15], Jun Ma[19], Xiaoping Yang[20], Elodie Puybareau[18], Ilkay Oksuz[17], Stephanie Bricq[16], Weisheng Li[15], Kumaradevan Punithakumar[14], Sotirios A. Tsaftaris[11], Laura M. Schreiber[10], Mingjing Yang[9], Guocai Liu[7,8], Yong Xia[6], Guotai Wang[5], Sergio Escalera[3,4], Xiahai Zhuang[1]*

[1] *School of Data Science, Fudan University, Shanghai, China*
[2] *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China*
[3] *Departament de Matemàtiques & Informàtica, Universitat de Barcelona, Barcelona, Spain*
[4] *Computer Vision Center, Universitat Autònoma de Barcelona, Spain*
[5] *School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[6] *School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China*
[7] *College of Electrical and Information Engineering, Hunan University, Changsha, China*
[8] *National Engineering Laboratory for Robot Visual Perception and Control Technology, Changsha, China*
[9] *College of Physics and Information Engineering, Fuzhou University, Fuzhou, China*
[10] *Chair of Molecular and Cellular Imaging, Comprehensive Heart Failure Center, Wuerzburg University Hospitals, Wuerzburg, Germany*
[11] *School of Engineering, University of Edinburgh, Edinburgh, UK*
[12] *School of Robotics, Xi'an Jiaotong-Liverpool University, Suzhou, China*
[13] *Department of Electrical and Computer Engineering, University of California, Los Angeles, USA*
[14] *Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton, Canada*
[15] *Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecomm-unications, Chongqing, China*
[16] *ImViA Laboratory, Université Bourgogne Franche-Comté, Dijon, France*
[17] *Computer Engineering Department, Istanbul Technical University, Istanbul, Turkey*
[18] *EPITA Research and Development Laboratory (LRDE), Le Kremlin-Bicêtre, France*
[19] *Department of Mathematics, Nanjing University of Science and Technology, Nanjing, China*
[20] *Department of Mathematics, Nanjing University, Nanjing, China*

1. Results of additional evaluation metrics

Table 1 presents the 95% Hausdorff distance (HD) and precision values of the evaluated algorithms. One can see that the best 95% HD value for edema was 6.42 ± 7.02 mm, which was worse than that for scars (4.79 ± 5.88 mm). This is due to the fact that edema generally has larger surface bordering with complex tissues, which could confuse the segmentation models. However, the average precision on edema was much better than that on scars (0.773 ± 0.156 vs 0.729 ± 0.290). It reveals the difficulty of identifying scars from noisy areas, which severely introduces false positives.

2. Correlation studies of clinical indices

To evaluate the performance of the benchmark methods on estimating the clinical indices, we performed correlation studies by comparing the prediction with the ground truth for each case. In total, we consider

*Senior and corresponding author:Xiahai Zhuang; †These authors contribute equally and are the co-first authors: Lei Li (lilei.sky@sjtu.edu.cn), Fuping Wu (17110690006@fudan.edu.cn) and Sihan Wang (shwang21@m.fudan.edu.cn).

Table 1: Summary of the quantitative evaluation results of scar and edema segmentation by the fifteen teams. Asterisk (*) indicates the method obtained statistically poorer results ($p < 0.01$) compared to the best performance in terms of corresponding metrics. HD: Hausdorff distance.

Team	Scar		Edema	
	95% HD (mm)	Precision	95% HD (mm)	Precision
UESTC	5.54 ± 7.80	0.705 ± 0.242	$\mathbf{6.42 \pm 7.02}$	0.754 ± 0.135
UBA	9.28 ± 10.6	0.696 ± 0.269	10.9 ± 8.38	0.707 ± 0.150
NPU	7.17 ± 10.9	0.668 ± 0.290	7.85 ± 9.44	0.740 ± 0.145
USTB	6.35 ± 8.47	$0.610 \pm 0.317^*$	8.69 ± 9.81	$0.670 \pm 0.156^*$
FZU	7.74 ± 7.96	0.641 ± 0.238	$12.1 \pm 11.0^*$	$0.624 \pm 0.156^*$
NJUST	6.78 ± 9.03	$\mathbf{0.729 \pm 0.290}$	6.92 ± 7.58	0.717 ± 0.137
CQUPT I	4.96 ± 7.49	0.720 ± 0.278	6.69 ± 8.87	$\mathbf{0.773 \pm 0.156}$
LRDE	8.43 ± 11.2	$0.578 \pm 0.271^*$	$12.6 \pm 9.95^*$	$0.624 \pm 0.129^*$
CQUPT II	5.19 ± 6.07	0.697 ± 0.257	7.70 ± 7.18	0.764 ± 0.109
HNU	9.28 ± 13.2	0.667 ± 0.276	9.12 ± 10.5	0.753 ± 0.110
Edin	6.75 ± 10.3	$0.628 \pm 0.320^*$	8.58 ± 8.50	$0.664 \pm 0.134^*$
UBO	9.83 ± 10.3	$0.572 \pm 0.328^*$	10.9 ± 10.3	0.669 ± 0.158
UHW	$\mathbf{4.79 \pm 5.88}$	0.669 ± 0.290	8.28 ± 6.86	$0.649 \pm 0.127^*$
ITU	8.87 ± 10.9	$0.626 \pm 0.295^*$	8.39 ± 7.44	0.698 ± 0.193
UOA	$10.5 \pm 11.7^*$	0.672 ± 0.287	10.4 ± 9.94	0.726 ± 0.151
Average	7.43 ± 9.62	0.659 ± 0.283	9.04 ± 8.94	0.702 ± 0.151

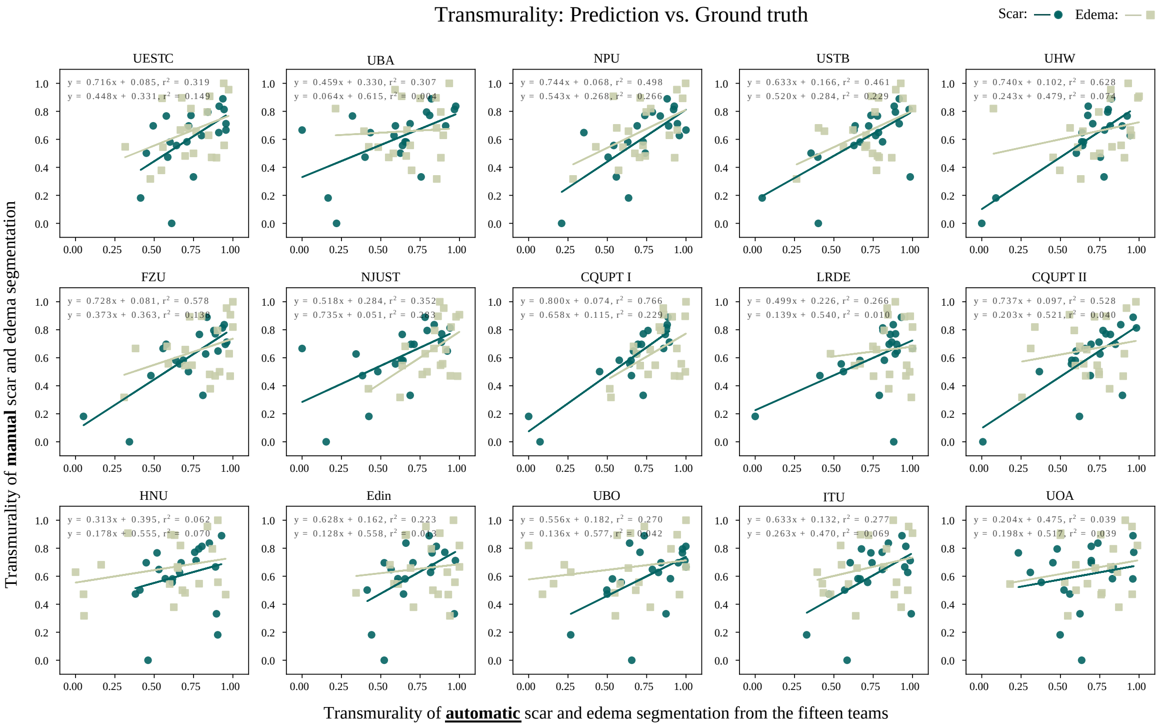


Figure 1: The scatter point plots and correlations between the transmurality of predicted and manual pathology segmentation.

three commonly used clinical indices, i.e., transmurality, surface area, and volume, for myocardial pathology analysis. Figure 1, Figure 2 and Figure 3 present the correlations of the three indices measured from automatic segmentation methods and manual segmentation. The correlation studies can be used as complements to the mean value based evaluation for two study groups (automatic vs. ground truth). For example, according to Table 8, for the clinical index "volume of scar", Edin has the best correlation with ground truth (in bold), but is statistically different compared to the mean of ground truth (marked with an asterisk).

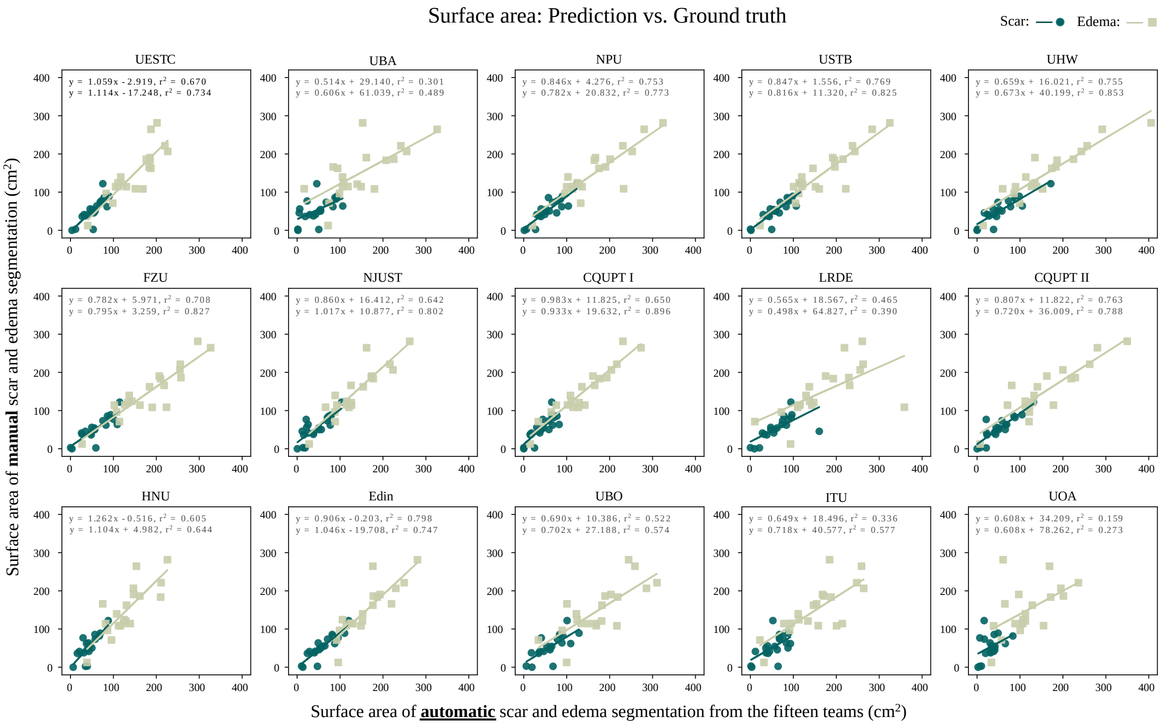


Figure 2: The scatter point plots and correlations between the surface area of predicted and manual pathology segmentation.

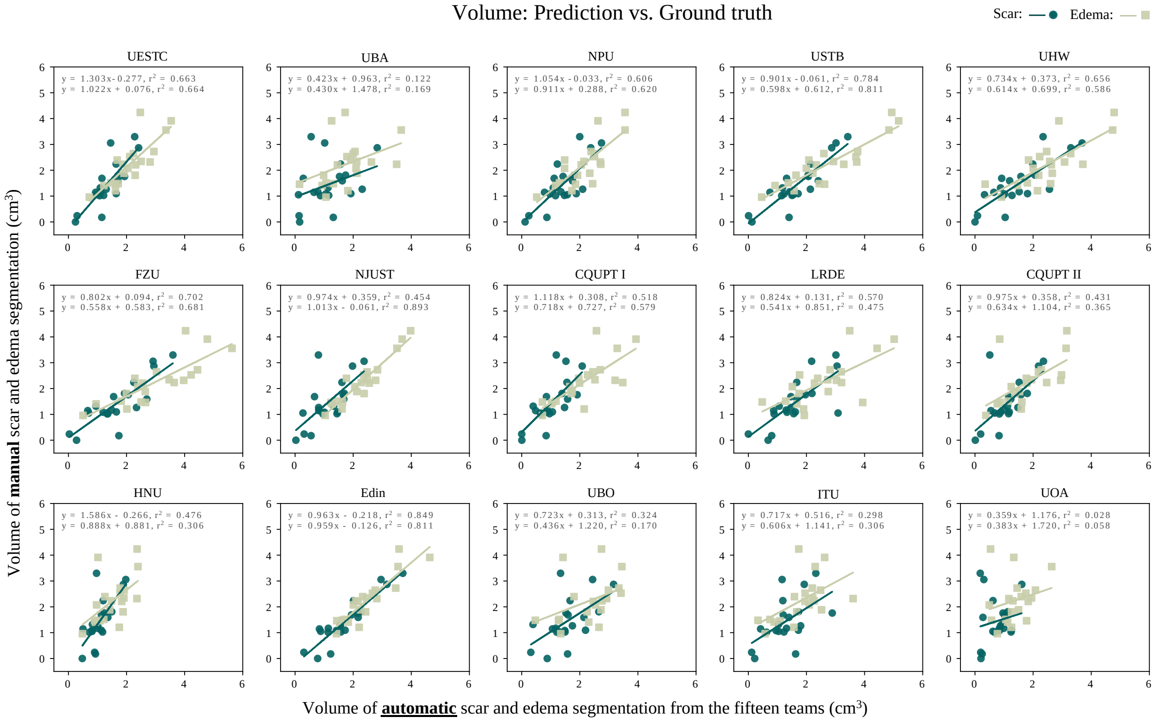


Figure 3: The scatter point plots and correlations between the volume of predicted and manual pathology segmentation.