Infarct Segmentation Challenge on Delayed Enhancement MRI of the Left Ventricle

Rashed Karim[1], Piet Claus[3], Zhong Chen[1,2], R. James Housden[1], Samantha Obom[1], Harminder Gill[1], YingLiang Ma[1], Prince Acheampong[1], Mark O'Neill[1,2], Reza Razavi[1,2], Tobias Schaeffter[1], and Kawal S. Rhode[1]

[1] Division of Imaging Sciences and Biomedical Engineering, King's College London, United Kingdom
[2] Department of Cardiology, Guy's and St. Thomas' NHS Foundation Trust, London, United Kingdom
[3] Cardiovascular Imaging and Dynamics, Department of Cardiovascular Sciences, KULeuven, Leuven, Belgium

Abstract. This paper presents collated results from the Delayed Enhancement MRI (DE-MRI) segmentation challenge at MICCAI 2012. DE-MRI Images from fifteen patients and fifteen pigs were randomly selected from two different imaging centres. Three independent sets of manual segmentations were obtained for each image and included in this study. A ground truth consensus segmentation based on all human rater segmentations was obtained using an Expectation-Maximization (EM) method (the STAPLE method). Automated segmentations from five groups contributed to this challenge.

Keywords: Segmentation, Delayed-Enhancement MRI, Left ventricle, Segmentation Challenge.

1 Introduction

In this era of timely access to the cardiac catheterization lab, treatment of myocardial infarction shifts from survival to reducing infarct size, post-reperfusion myocardial vascular obstruction (MVO) and hemorrhage. In this setting, Contrast-Enhanced Magnetic Resonance Imaging (CE-MRI) has become an indispensable imaging modality to assess MVO, usually imaged a few minutes after contrast administration and also infarct size, assessed after 15 minutes and hence referred to as Delayed Enhancement MRI (DE-MRI) [1]. In the last decade infarct size assessed with DE-MRI has been increasingly used as a primary end-point in clinical studies. Moreover, interventions target the infarct border zone, and the use of fluoroscopy augmented with MRI-based anatomical models of the heart and infarct areas are entering the clinical work-flow (e.g. for VT ablation or localized delivery in cell-based therapies). These developments require both a relatively fast and accurate segmentation of the infarct region.

Manual delineation of the enhanced myocardial regions is relatively time consuming and requires training. Though expert consensus on delineation of myocardial infarction on DE-MRI can be reached, manual delineations still suffer

O. Camara et al. (Eds.): STACOM 2012, LNCS 7746, pp. 97–104, 2013.

from inter- and intra-observer variability. Automated thresholding based on the full-width half maximum signal intensity or a number of standard deviations above normal myocardial signal intensity, remain acquisition dependent.

In this challenge, DE-MRI images of left ventricles containing infarct regions, from both human and porcine studies, were provided to the participating groups along with a given myocardial segmentation. To minimize inter-observer variability in the ground-truth infarct segmentations, a probabilistic estimate was computed for each dataset from three different expert delineations.

2 Methods

2.1 MRI Data

Cardiac DE-MRI images were collected at two centers (King's College London (KCL) and Katholieke Universiteit Leuven (KUL)), providing fifteen human and fifteen porcine datasets. For all datasets a short axis stack of DE-MRI images covering the left ventricle were provided together with a mask of the left-ventricular myocardium. A sample of the datasets can be seen in Fig. 1.

The human datasets were randomly selected patients with a known history of ischaemic cardiomyopathy and under assessment for implantable cardioverter defibrillator (ICD) for primary or secondary preventions after infarction. In addition to this, the patients chosen had a history of myocardial infarction at least 3 months prior to their MRI scan along with evidence of significant coronary artery disease on angiography and evidence of left ventricular impaired systolic function on echocardiography. The images were acquired on a clinical 1.5T MRI unit (Achieva, Philips, The Netherlands)

The fifteen porcine studies were randomly selected from an experimental database of a pre-clinical model of chronic myocardial ischemia [2], with either left-anterior descending or left-circumflex artery induced lesions. Datasets were acquired six weeks after the induction of the coronary lesion on a clinical 3T MRI unit (Trio, Siemens, Erlangen, Germany). Details of the acquisition can be found in Table 1.

2.2 Raters

There were five automated raters or algorithms (UPF, MCG, MVS, KCL, ALM) and three expert human raters (HA, HB, HC) taking part in this study. Brief descriptions of the algorithms are given in Table 2.

2.3 Evaluation

The accuracy and performance of each algorithm was evaluated by comparing its segmentations against the human raters. For rater segmentations, the Simultaneous Truth And Performance Level Estimation (STAPLE) method [3] was used to obtain a single ground truth. STAPLE estimates the ground truth by forming

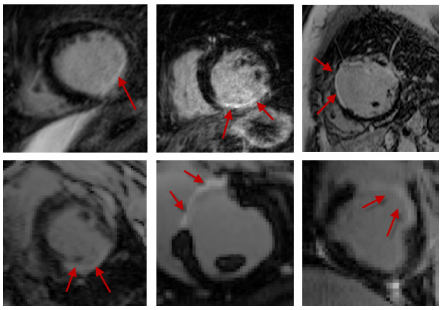

Fig. 1. Sample of the human (top row) and animal (bottom row) DE-MRI data provided to participants as part of the challenge. Red arrows indicate areas of enhancement.

Table 1. Image acquisition

	KCL	**KUL**
Scanner type	Philips Achieva 1.5T	Siemens Trio 3.0T
Sequence	Segmented 2D, Inversion recovery Gradient Echo ECG triggered, breathold	Segmented 3D Inversion recovery Gradient Echo ECG triggered, breathold
TI, TR, TE, FA	280 ms, 3.4 ms, 2.0 ms, 25°	340-370 ms, 2.19 ms, 0.78 ms, 15°
Voxel size	$1.8 \times 1.8 \times 8$ mm	$1.8 \times 1.8 \times 6$ mm
Interleaving	Every RR	Every other RR

Image acquisition parameters for the challenge DE-MRI data. Abbreviations: TI - Inversion time, TR - Repetition time, TE - Echo time, FA - Flip angle.

an optimal combination of the segmentations, by weighting each segmentation depending upon the estimated performance level, together with a prior model that can account for the spatial distribution of structures and spatial homogeneity constraint. In this work, the threshold set on the probabilities obtained from STAPLE was 0.7 and above for a pixel to be labeled as scar.

In addition to computing STAPLE of human rater segmentations, the STAPLE of all submissions was also computed with a leave-one-out approach. For example, UPF was tested against the STAPLE of KCL, MVS, MCG, ALM combined. This assessed how the submissions fared between themselves. To assess similarity between segmentations the Dice co-efficient was used [4]:

$$\mathcal{D} = \frac{2|X \cap T|}{|X| \cup |T|} \tag{1}$$

where X and T are sets of pixels belonging to the algorithm's output and ground truth respectively. The Dice is normalized to 100 for convenience. Furthermore,

Table 2. Algorithms presented at the challenge

Algo.	Description	Auto or Semi-auto
UPF	Region growing and morphology	Auto
MCG	Conditional Random Fields (CRF)	Auto
MV	Gaussian mixture, EM-algorithm, Watershed transformation	Auto
KCL	Markov Random Fields, Graph-cuts	Auto
ALM	Support Vector Machines followed with Level-set evolution	Semi-auto

Institution abbreviations: UPF - Universitat Pompeu Fabra, MCG - McGill University, MV - Mevis Fraunhofer, KCL - King's College London, ALM - Alma IT Systems.

the amount of infarct detected (V_{IF}) represented as percentage of myocardium V_{MYO} was determined in all methods as:

$$\mathrm{IF}\% = \frac{V_{\mathrm{IF}}}{V_{\mathrm{MYO}}} \times 100\% \tag{2}$$

3 Results

Following the submission of results from each group of the challenge, segmentations were evaluated against the human rater segmentations using Dice and infarct volumes. The human and porcine images were analyzed separately.

3.1 Human Patient Datasets

Segmentations were compared to ground truth and between themselves with a leave-one-out approach (see example dataset in Fig. 2). Dice results can be seen in Fig. 3. To measure how much the infarct volumes (expressed as a percentage of total myocardium) differ from the human raters' segmentations, Bland-Altman plots are presented in Fig. 4.

3.2 Porcine Datasets

Similar to the patient datasets, segmentations were compared between ground truth and with the leave-one-out approach. An example on a dataset can be found in Fig. 5. The Dice comparison is given in Fig. 6. Difference in percentage of infarct volumes is shown in the Bland-Altman plots of Fig. 7.

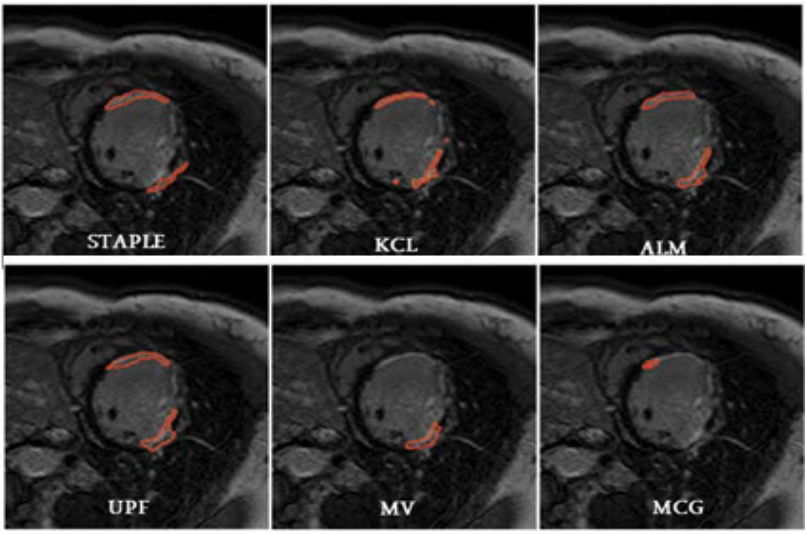

Fig. 2. Human Dataset: comparing STAPLE with submissions in a single slice of an example dataset

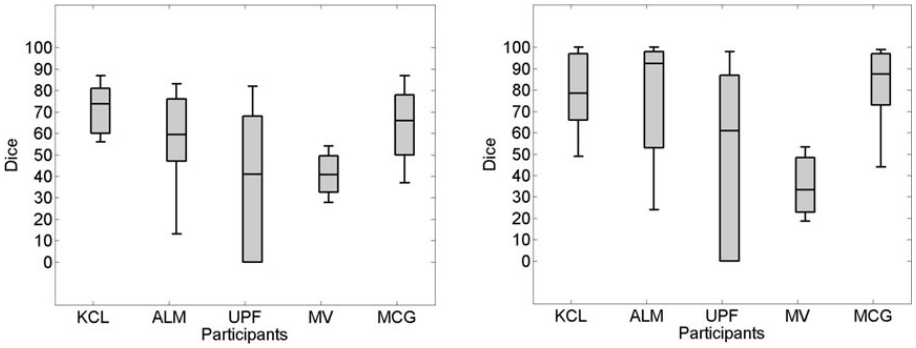

Fig. 3. Human Datasets: comparing segmentations from each submission using Dice - (Left) against STAPLE ground-truth, and (Right) against STAPLE of leave-one-out for each submission. For example, UPF here is tested against the STAPLE of KCL, MVS, MCG, ALM. Each box in the plot represents lower, middle and upper quartiles.

3.3 Discussion

A collation study for the DE-MRI segmentation of enhanced tissue representing fibrosis and scar has been presented in this work. A ground-truth segmentation was generated using the STAPLE algorithm combining three human rater segmentations. The performance of each algorithm taking part in the study was compared to the STAPLE estimate of the ground truth. Each algorithm's output was also compared to segmentations from other algorithms using a leave-one-out

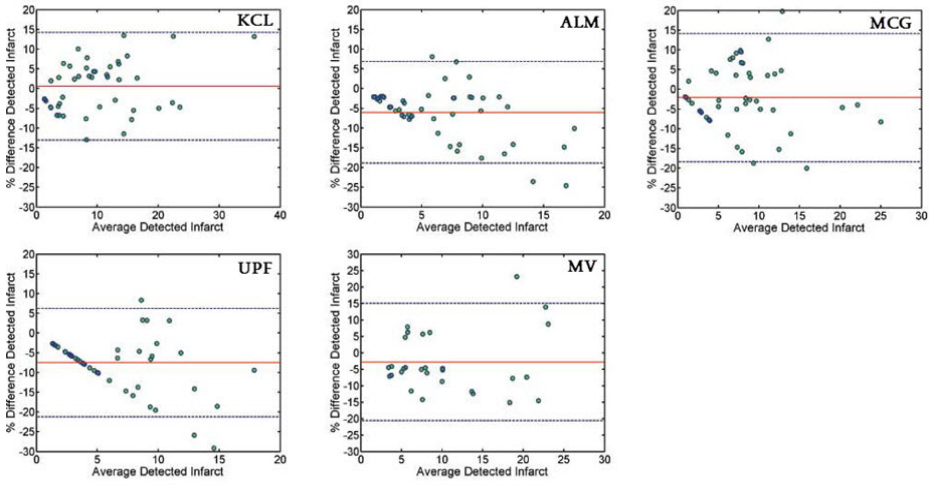

Fig. 4. Human Datasets: Bland-Altman plots showing differences in infarct volumes (%) in segmentations of each submission against STAPLE of human raters. The solid line shows the mean and the dashed lines show ±2 standard deviations.

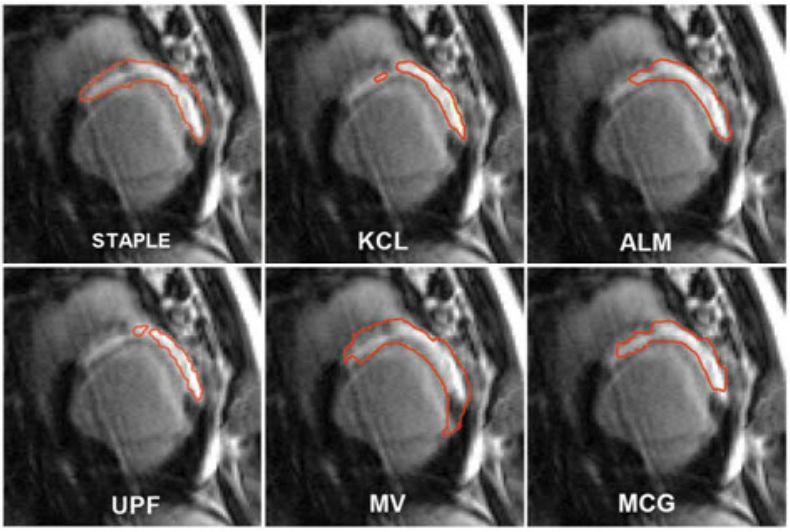

Fig. 5. Porcine Dataset: comparing STAPLE with submissions in a single slice of an example dataset

STAPLE result. The STAPLE method is able to resolve disagreement between human raters especially in regions where it is difficult to assert on infarction. In the ventricle, this is especially within the left ventricular outflow tract (LVOT) where fibrotic tissue making up the valve is not infarcted. Furthermore, using

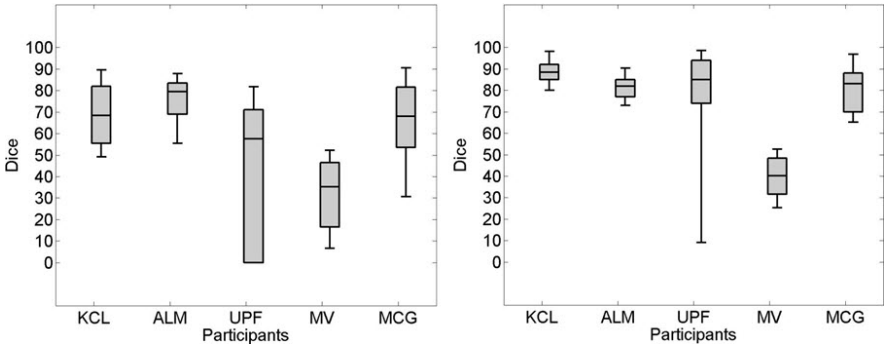

Fig. 6. Porcine Datasets: comparing segmentations from each submission using Dice - (Left) against STAPLE ground-truth, and (Right) against STAPLE of leave-one-out for each submission. Each box in the plot represents lower, middle and upper quartiles.

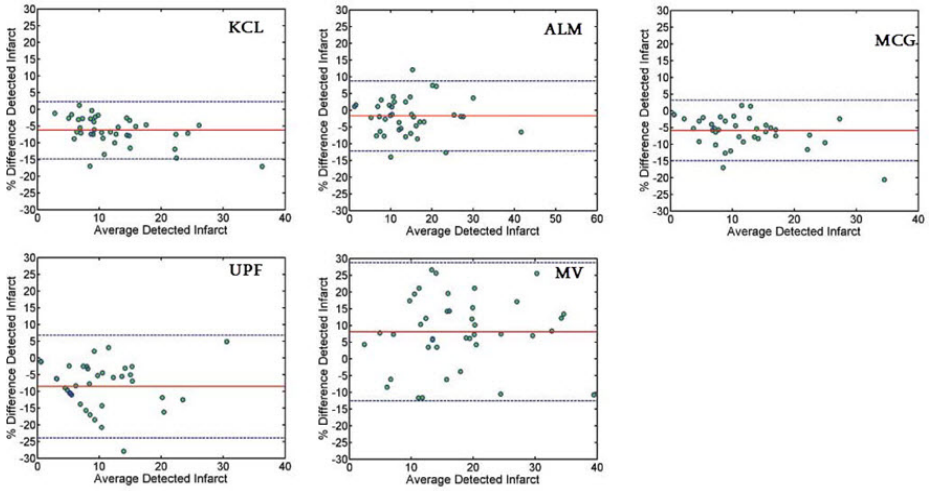

Fig. 7. Porcine Datasets: Bland-Altman plots showing differences in infarct volumes (%) in segmentations of each submission against STAPLE of human raters. The solid line shows the mean and the dashed lines show ±2 standard deviations.

the STAPLE result of the algorithms' outputs (i.e. leave-one-out approach), each algorithm's performance can be measured against the others and this reveals interesting insights into whether certain algorithms are computing scar in a similar fashion.

It was observed that in general, porcine scans produced better segmentations compared to human scans. One major contributing factor is the quality of porcine scans, providing excellent contrast for infarcted regions. In human datasets, it is often challenging to obtain good contrast due to the nature of the scan. For example, incorrect selection of inversion time causes myocardium

not to be nulled properly and if the patient cannot hold their breath during the scan it produces breathing artefacts. For patient scans, obtaining a Dice of 80 or above was challenging for all submissions. When considering infarct volumes, submissions differed by a maximum of approximately ±15% and ±10% for patient and porcine scans respectively. This further shows that segmenting patient scans was far more challenging.

The Dice used in this study has several limitations. It can be over-sensitive to a small mis-match. An alternative approach for comparison is thus considered. Comparing infarct using percentage volumes has become a simple and standard approach. Infarct volumes are computed on selected slices and represented in the difference plots of Figs. 4 and 7. Some patterns are notable, for example the data points which appear on a straight line in Fig. 4 (MCG and UPF) and Fig. 7 (UPF). This is due to the submission recording a zero volume for each of those data points and thus selected slice, resulting in a straight-line on the difference-vs-average plot.

A second limitation is the inter-observer variation of human rater segmentations. There was less variation in the porcine datasets than the human datasets. All datasets were segmented by individuals with experience working on these scans. They were segmented using ITK-SNAP (www.itk-snap.org) [5] which is a commonly used tool for performing manual segmentations.

3.4 Conclusions

The presented work reports on the preliminary results of the ventricle DE-MRI segmentation challenge at the MICCAI 2012 meeting. Future work will look to provide a more extensive analysis of the submitted results.

Acknowledgements The authors would like to thank the members of the Division of Biomedical Engineering and Imaging Sciences, King's College London, who assisted with this study. This work was funded by a research grant from the Medical Engineering Council (MEC), UK.

References

1. Kellman, P., Arai, A.: Cardiac imaging techniques for physicians: Late enhancement. Journal of Magnetic Resonance Imaging 36(3), 529–542 (2012)
2. Wu, M., D'hooge, J., Ganame, J., Ferferieva, V., Sipido, K., Maes, F., Dymarkowski, S., Bogaert, J., Rademakers, F., Claus, P.: Non-invasive characterization of the area-at-risk using magnetic resonance imaging in chronic ischaemia. Cardiovascular Research 89(1), 166–174 (2011)
3. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23(7), 903–921 (2004)
4. Dice, L.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945)
5. Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G.: User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31(3), 1116–1128 (2006)