Published in: European Journal of Radiology

doi.org:10.1016/j.ejrad.2021.109817

Self-configuring nnU-net pipeline enables fully automatic infarct segmentation in late enhancement MRI after myocardial infarction

Julius F. Heidenreich[a*], Tobias Gassenmaier[a], Markus J. Ankenbrand[b,c], Thorsten A. Bley[a], Tobias Wech[a]

[a] Department of Diagnostic and Interventional Radiology, University Hospital Würzburg, Germany

[b] Department of Cellular and Molecular Imaging, Comprehensive Heart Failure Center, University Hospital Würzburg, Germany

[c] Center for Computational and Theoretical Biology, University of Würzburg, Germany

*Corresponding author: heidenreic_j@ukw.de

Highlights

- A 2D nnU-net trained for segmentation of myocardial scar tissue in late enhancement MRI using magnitude and PSIR reconstructions as dual contrast input predicted labels of myocardium with a Dice similarity coefficient of 0.83 ± 0.03 and of scar tissue with 0.72 ± 0.08.
- No significant difference was observed when comparing the performance of trained 2D nnU-net using a single contrast input (magnitude or PSIR) or a dual contrast input (magnitude and PSIR).
- The nnU-net pipeline enables the training of a network with a high performance without the necessity for manual configuration of the network design and the choice of hyperparameters.

Abstract

Purpose

To fully automatically derive quantitative parameters from late gadolinium enhancement (LGE) cardiac MR (CMR) in patients with myocardial infarction and to investigate if phase sensitive or magnitude reconstructions or a combination of both results in best segmentation accuracy.

Methods

In this retrospective single center study, a convolutional neural network with a U-Net architecture with a self-configuring framework ("nnU-net") was trained for segmentation of left ventricular myocardium and infarct zone in LGE-CMR. A database of 170 examinations from 78 patients with history of myocardial infarction was assembled. Separate fitting of the model was performed, using phase sensitive inversion recovery, the magnitude reconstruction or both contrasts as input channels.

Manual labelling served as ground truth. In a subset of 10 patients, the performance of the trained models was evaluated and quantitatively compared by determination of the Sørensen-Dice similarity coefficient (DSC) and volumes of the infarct zone compared with the manual ground truth using Pearson's r correlation and Bland-Altman analysis.

Results

The model achieved high similarity coefficients for myocardium and scar tissue. No significant difference was observed between using PSIR, magnitude reconstruction or both contrasts as input (PSIR and MAG; mean DSC: 0.83 ± 0.03 for myocardium and 0.72 ± 0.08 for scars). A strong correlation for volumes of infarct zone was observed between manual and model-based approach (r = 0.96), with a significant underestimation of the volumes obtained from the neural network.

Conclusion

The self-configuring nnU-net achieves predictions with strong agreement compared to manual segmentation, proving the potential as a promising tool to provide fully automatic quantitative evaluation of LGE-CMR.

Abbreviations

2D - two dimensional

3D - three dimensional

CMR - cardiac magnetic resonance imaging

DSC - Sørensen-Dice similarity coefficient

LGE - late gadolinium enhancement

MAG - magnitude reconstruction

MI - myocardial infarction

NN - neural network

PSIR - phase sensitive inversion recovery

Key Words

Deep Learning, CMR, Segmentation, Myocardial Infarction, Scar, nnU-net

Introduction

During acute and chronic myocardial infarction (MI), detection and quantification of non-viable myocardium is evolving as biomarker for disease management and treatment planning, as it can be used as predictor for disease severity and for disease monitoring [1,2]. Cardiovascular magnetic resonance (CMR) with late gadolinium enhancement (LGE) has been established as reference technique for the measurement of the volume of viable and non-viable tissue [3,4]. For qualitative and reproducible evaluation of CMR, recommendations for a standardized post-processing were proposed by the Task Force for Post Processing of the Society for Cardiovascular MR (SCMR), including dedicated hard- and software as well as instructions for a standardized manual post-processing by experienced readers [5].

Neural networks, specifically trained with pre-labeled datasets, have evolved into promising approaches for fully automatic image segmentation [6–8] . In particular, a plethora of studies has proven the power of convolutional neural networks in pixel classification and semantic segmentation of myocardial tissue in recent years [9–15]. For the segmentation of myocardial infarct zones within LGE MR images, Fahmy et al. [11] trained a 2D U-net architecture using a cohort of patients with hypertrophic cardiomyopathy. Inference on test data resulted in a mean Sørensen-Dice similarity coefficient (DSC) of 0.58. If the myocardium was manually segmented by a human operator as a pre-step, the performance of a NN-based myocardial scar segmentation could be increased to a median DSC of 0.71 in Moccia et al. [10]. While the latter approach still requires cumbersome pre-processing, DSC of the fully automatic method presented by Fahmy et al. was rather low.

One general issue of all cited deep learning approaches is the complex choice of the exact design and configuration of the network architecture as well as the according training. While the subsequent application of the networks might provide highly automatic and fast semantic image segmentation, the optimization process is typically manual and not seldomly

represented by a time-consuming heuristic search for the best setup. Even worse, this frequently results in sub-optimal solutions, not unleashing the full potential of the generic model [16].

In Isensee et al. [17], a self-configuring solution for deep learning based semantic segmentation was introduced, which is condensing general rules e.g. for detailed network design and hyperparameter choices from a broad spectrum of empirical observations. As no new architecture, but rather an automated procedure in terms of a robust end-to-end deep learning pipeline was presented, the underlying network was dubbed nnU-net ('no new net'). The authors reported on superior performances with respect to various specialized deep learning solutions for medical imaging, ranging from the segmentation of liver tumors, to cell tracking to manifold types of post processing in cardiac imaging.

In this paper, we trained and evaluated a 2D version of nnU-net for fully automatic segmentation of infarct zones in LGE MR images, using MRI data in phase sensitive reconstruction (phase sensitive inversion recovery, PSIR [18]) and magnitude reconstruction from a cohort of patients with history of myocardial infarction. nnU-net was trained with both contrasts (PSIR and magnitude reconstruction) separately, as well as with a combined version using both contrasts as separate input channels. Performances of the obtained models were ultimately evaluated using a test dataset, previously unseen by nnU-net.

Methods

The data were collected and analyzed in accordance with the relevant guidelines and regulations. All methods of this study were performed in accordance with the Declaration of Helsinki. The study was approved by the institutional review board of the University of Würzburg and the requirement of a written informed consent by the study participants was waived due to the retrospective study design (IRB code: DE/EKBY13, decision number: 20200324-01). The data was fully anonymized for data analysis.

Image database and data curation

In this retrospective single-center study we assembled a database of standardized contrast-enhanced CMR examinations between 2015 and 2018 on a 3T MR system (MAGNETOM Prisma, Siemens Healthcare, Erlangen, Germany) from patients with the diagnosis of myocardial infarction. Inclusion criterium was the known history of myocardial infarction. Patients with acute and chronic myocardial infarction were included. The timepoints of MR imaging were 0-5 days after infarction (acute MI), 7 days after infarction (subacute MI) up to 12 months after infarction (chronic MI). Exclusion criteria were extensive artifacts as well as incomplete datasets due to interrupted image acquisition. The finally assembled database of 170 examinations originate from 78 patients (65 men, mean age 64.3 years, age range 35 – 87 years). 61 datasets were acquired in the setting of acute MI, 59 datasets were acquired during subacute MI and 50 datasets were acquired after 12 months in the setting of chronic MI. Further demographics are summarized in Table 1.

LGE images in short axis orientation were acquired from base to apex using a 2D fast low-angle shot inversion recovery sequence 10 min after intravenous injection of a Gd-based contrast agent (DOTA-Gd). Both PSIR and magnitude reconstructions (MAG) of the images as well as a dual contrast version (PSIR and MAG as separate channels, referred to as DUAL) were stored in NIfTI format. Thereby, images were cropped to a uniform size of 128 × 128

pixels. The field of view was chosen such that a generous distance to the heart was ensured for typical planning of short axis orientations. All segmentation labels were resized accordingly.

The database of 170 examinations (1902 images) was separated patient wise into 159 examinations for training and cross-validation (1806 images, 68 patients) and 11 examinations for testing (96 images, 10 patients). There was no interference between training and testing data. Repeated examinations in a single patient were holistically attributed to one of the two groups. For improved class balancing, edge slices without infarct zones were excluded from training data. In addition, complementing images acquired outside the left ventricle, images afflicted by artifacts and slices showing a large number of flying pixels in the manual segmentation masks (see next section) were removed. This resulted in a total number of 885 images for training with present infarct in 699 images (79 %) and 96 images for testing with present infarct in 72 images (75 %; Figure 1).

Definition of Ground Truth

For the definition of the myocardium, the endo- and epicardial contours were delineated manually by an experienced operator (x.x., board certified radiologist with 8 years of experience in CMR) using cvi42 (Circle Cardiovascular Imaging Inc., Calgary, Canada) and the original images in DICOM format. Additionally, a reference contour of at least 4 x 4 pixels was delineated for the definition of remote myocardium (myocardium without visual late gadolinium enhancement at the opposite site of infarct zone)[19]. For quantification of infarct zone, enhanced areas were obtained numerically within the magnitude reconstruction employing the n+5·SD approach as recommended by the Society for Cardiovascular Magnetic Resonance (SCMR) [5]. In this technique, pixels with a signal intensity which is higher by 5 times of the standard deviation compared with remote myocardium are defined as infarct zone. Segmented volumes were reviewed by the experienced operator and corrected, if needed. The according PSIR reconstructions were included in this review process. Papillary

muscles in the left ventricle were excluded from segmentation. Areas of microvascular obstruction were added manually to the area of infarction, yielding final labels with a pixel classification for background, myocardium and infarct zone.

Neural network

A 2D nnU-net was set up and trained in python as described in the paper [17] and its accompanying online repository [20]. Its fixed design relies on plain convolutions (conv), instance normalization (norm) and leaky rectified linear units (ReLU). Resolution is reduced after two blocks consisting of conv – normalization – ReLU by means of strided convolutions. In the upsampling path transposed convolutions are employed.

The network architecture is adapted e.g. with respect to its depth by means of fixed rules on information gathered from data fingerprints. Further rule-based adjustments are the batch and the patch size, data normalization and resampling choices.

Training is performed in 5-fold cross-validation with 1000 epochs for each run. Stochastic gradient descent is used for optimization with an initial learning rate of 0.01, continuously decreased close to zero towards the final epoch. A combination of dice loss and cross entropy loss is used as cost measure. To reduce oversampling, a range of data augmentation transformation is applied including scaling, rotation, addition of Gaussian noise and blurring.

More details of the fixed, rule-based and empirical parameter choices of the nnU-net pipeline can be found in [17] and its supplements.

The trainings of the different models in this work were performed on two Nvidia Titan XP and two Nvidia Titan RTX GPUs.

Evaluation

The models which were trained separately for PSIR, MAG and DUAL were applied to the corresponding PSIR, MAG and DUAL images of the previously unseen 11 patients. The obtained areas for the two classes of myocardium and infarct zone were compared to the ground truth of manual segmentations by calculating the DSC. LGE areas smaller than three

pixels were set to zero in this evaluation. Analysis of variances (ANOVA) was used to test for significant differences.

Furthermore, the volume of infarct zones was determined and compared between the automatic deep-learning-based and the reference manual approach by means of a Bland-Altman plot.

MATLAB (v2020a, TheMathWorks, Natick, USA) was used for all analysis.

Results

Self-configuration of the nnU-net pipeline

The preprocessing of a data fingerprint by the nnU-net pipeline resulted in a training patch size of 128 × 128 and a batch size of 44 for all groups (PSIR, MAG, DUAL). The trained network architecture has six resolution stages, convolution kernel sizes were [[3, 3], [3, 3], [3, 3], [3, 3], [3, 3], [3, 3]] and downsampling strides [[2, 2], [2, 2], [2, 2], [2, 2], [2, 2]]. Automatic segmentation of one 2D image took 0.1 s using the trained nnU-net on a Nvidia Titan RTX GPU.

Comparison of the model's predictions to the ground truth

Obtained DSC values are listed in Table 2 as means ± standard deviations. Means are in general higher when calculated across patients (11 3D-samples) as compared to the evaluation per-slice (97 2D-samples). The values differ only slightly between the different groups of PSIR, MAG and DUAL, indicating similar performance of the automatic segmentation. This was confirmed by statistical testing of DSC, which did not indicate a significant difference ($p>0.05$) between the three groups, neither for myocardial nor for infarct size (both slice- and patient-wise evaluation). Exemplary images comparing the automatic segmentation performed by the nnU-net with manual ground truth are depicted in Figure 2.

Volumetry of infarct zone

Comparison of the infarct volumes between nnU-net derived and manual segmentations resulted in high Pearson correlation coefficients of $r = 0.96$ for the PSIR approach, $r = 0.92$ for MAG and $r = 0.96$ for DUAL. nnU-net underestimates infarct volumes with a bias (difference of means with respect to the mean of the ground truth) of 23%, 24% and 23% for PSIR, MAG and DUAL, respectively (Figure 3). The drifts are throughout significant ($p < 0.05$); corresponding standard deviations for the differences appear comparatively low (8.5 cm^3, 8.1 cm^3 and 7.3 cm^3).

Discussion

In this study a deep learning approach based on a two-dimensional neural network with a U-net architecture and a self-configuring training pipeline (nnU-net) was evaluated for fully automatic segmentation of infarct zones in late gadolinium enhanced cardiac MRI. The network was trained using a database of 885 images from 68 patients (159 examinations) and subsequently tested in 10 patients (97 images; 11 examinations). Three networks were investigated separately, corresponding to phase sensitive (PSIR) and magnitude reconstructions (MAG) of the inversion recovery acquisition, as well as one combined group with both contrasts as input (DUAL).

The achieved DSC values for the test dataset were not significantly different for the three contrast groups, suggesting that the choice between the two LGE approaches might not be decisive for the subsequent evaluation and quantification. It is noteworthy, however, that the results might be biased by the way ‘ground truth’ is defined in our dataset. Both MAG and PSIR images were used for determining the initial manual segmentation: Intensity-based automatic labelling of infarct zones was performed on the MAG depiction by use of the n+5·SD approach, while PSIR images were used as supplemental in the final supervision and refinement of segments by the expert radiologist. These manual segmentation labels were then used for all three groups, both for training and testing. Even though not practical, histological correlates would be necessary for real ‘ground truth’ and labels would need to be determined fully independently for the three groups. While the latter point is not entirely theoretical it comes with an extreme additional expenditure of manual segmentation time. Nevertheless, the result proves that the initial information gathered in both contrasts by the expert could be extracted from each individual alone by the neural network.

The segmentation performance for infarct zones as measured by DSC was overall high for the nnU-net approach as compared to the fully automated method presented by Fahmy et al [11],

who reported a DSC per slice of 0.58 ± 0.28 and per patient of 0.57 ± 0.23. According values in our study were 0.67 ± 0.24 (per-slice) and 0.72 ± 0.08 (per-patient) for the DUAL approach (other groups similar and not significantly different). These values obtained by the fully automated procedure correspond well with the performance achieved by the semi-automatic approach published by Moccia et al.[10] (median DSC 0.71, inter-quartile-range 0.32), which, however, comes with significant manual operator time for data pre-processing.

Infarct volumes showed overall strong correlation between automatic determination and manual reference. For the mean absolute values, however, a significant underestimation for the nnU-net was observed. A bias of similar extent was already observed in the study of Fahmy et al.[11], where the presented convolutional network predicted a infarct volume of 6.1 ± 7.4 cm^3 compared to a volume of 7.5 ± 8.5 cm^3 from the manual reference. While in[11] no loss function was specified, the nnU-net used in our study partly relies on Dice loss, which is known to potentially introduce a volume bias in the case of uncertainty[21], which is certainly given for the very challenging task of infarct segmentation. A study of various loss functions and their effect on the final volume was, however, beyond the scope of our work.

A limitation of the presented study is the size and scope of our data for training and testing the neural network. Diagnostics of LGE imaging in clinical routine is predominantly of qualitative nature such that segmentation labels are typically scarce. As only data from our institute was applied, generalization for other scanners and sites needs to be proven in further studies. With more data, application of 3D models as e.g. included in the nnU-net framework become meaningful and promising. Furthermore, the network was trained using a database comprising patients in acute and chronic state of myocardial infarction; the pixel classification layer and the softmax were trained to segment areas of late gadolinium enhancement, but not to differentiate acute or chronic MI. It would be tempting to add another classifier to differentiate between acute and chronic infarction as presented by Larroza et al. using machine learning driven texture analysis[22].

In conclusion, the presented model for automatic segmentation of infarct zones in late gadolinium enhancement cardiovascular MRI showed good performance without the need of any image pre-procession and pre-definition of region of interest and with higher mean DSC as presented in earlier fully automatic deep learning-based approaches. No significant differences with respect to DSC was found between the typical contrasts of magnitude reconstruction and PSIR and neither with respect to a joint input using both contrasts. These promising results suggest deep learning algorithms to replace the cumbersome and operator-biased manual segmentation of late gadolinium enhancement CMR images in near future.

References

1. Kim, H. W., Farzaneh-Far, A. & Kim, R. J. Cardiovascular Magnetic Resonance in Patients With Myocardial Infarction. Current and Emerging Applications. *Journal of the American College of Cardiology* (2009). doi:10.1016/j.jacc.2009.06.059
2. Kelle, S. *et al.* Prognostic Value of Myocardial Infarct Size and Contractile Reserve Using Magnetic Resonance Imaging. *J. Am. Coll. Cardiol.* (2009). doi:10.1016/j.jacc.2009.07.027
3. Kim, R. J. *et al.* Performance of delayed-enhancement magnetic resonance imaging with gadoversetamide contrast for the detection and assessment of myocardial infarction : An international, multicenter, double-blinded, randomized trial. *Circulation* (2008). doi:10.1161/CIRCULATIONAHA.107.723262
4. Baron, N. *et al.* Comparison of various methods for quantitative evaluation of myocardial infarct volume from magnetic resonance delayed enhancement data. *Int. J. Cardiol.* (2013). doi:10.1016/j.ijcard.2012.03.056
5. Schulz-Menger, J. *et al.* Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) Board of Trustees Task Force on Standardized Post Processing. *J. Cardiovasc. Magn. Reson.* (2013). doi:10.1186/1532-429X-15-35
6. Chen, C. *et al.* Deep learning for cardiac image segmentation: A review. *arXiv* (2019). doi:10.3389/fcvm.2020.00025
7. Moeskops, P. *et al.* Automatic Segmentation of MR Brain Images with a Convolutional Neural Network. *IEEE Trans. Med. Imaging* **35**, (2016).
8. Asgari Taghanaki, S., Abhishek, K., Cohen, J. P., Cohen-Adad, J. & Hamarneh, G. Deep semantic segmentation of natural and medical images: a review. *Artif. Intell. Rev.* (2020). doi:10.1007/s10462-020-09854-1

9. Zhang, N. *et al.* Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI. *Radiology* (2019). doi:10.1148/radiol.2019182304

10. Moccia, S. *et al.* Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images. *Magn. Reson. Mater. Physics, Biol. Med.* **32**, 187–195 (2019).

11. Fahmy, A. S. *et al.* Automated Cardiac MR Scar Quantification in Hypertrophic Cardiomyopathy Using Deep Convolutional Neural Networks. *JACC: Cardiovascular Imaging* (2018). doi:10.1016/j.jcmg.2018.04.030

12. Chen, C. *et al.* Deep Learning for Cardiac Image Segmentation: A Review. *Front. Cardiovasc. Med.* **7**, (2020).

13. Bai, W. *et al.* Automated cardiovascular magnetic resonance image analysis with fully convolutional networks 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing. *J. Cardiovasc. Magn. Reson.* **20**, (2018).

14. Bartoli, A. *et al.* Deep Learning–based Automated Segmentation of the Left Ventricular Trabeculations and Myocardium on Cardiac MR Images: A Feasibility Study. *Radiol. Artif. Intell.* (2020). doi:10.1148/ryai.2020200021

15. Zhu, Y., Fahmy, A. S., Duan, C., Nakamori, S. & Nezafat, R. Automated Myocardial T2 and Extracellular Volume Quantification in Cardiac MRI Using Transfer Learning–based Myocardium Segmentation. *Radiol. Artif. Intell.* **2**, (2020).

16. Waring, J., Lindvall, C. & Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine* **104**, (2020).

17. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* (2020). doi:10.1038/s41592-020-01008-z

18. Kellman, P., Arai, A. E., McVeigh, E. R. & Aletras, A. H. Phase-sensitive inversion

recovery for detecting myocardial infarction using gadolinium-delayed hyperenhancement. *Magn. Reson. Med.* **47**, (2002).

19. Bulluck, H., Dharmakumar, R., Arai, A. E., Berry, C. & Hausenloy, D. J. Cardiovascular Magnetic Resonance in Acute ST-Segment–Elevation Myocardial Infarction. *Circulation* **137**, 1949–1964 (2018).
20. DKFZ. nnUNet repository. *GitHub - downloaded 20.01.2021* https://github.com/MIC-DKFZ/nnUNet (2021). doi:https://github.com/MIC-DKFZ/nnUNet
21. Bertels, J., Robben, D., Vandermeulen, D. & Suetens, P. Theoretical analysis and experimental validation of volume bias of soft Dice optimized segmentation maps in the context of inherent uncertainty. *Med. Image Anal.* **67**, (2021).
22. Larroza, A. *et al.* Differentiation between acute and chronic myocardial infarction by means of texture analysis of late gadolinium enhancement and cine cardiac magnetic resonance imaging. *Eur. J. Radiol.* **92**, 78–83 (2017).

Figures

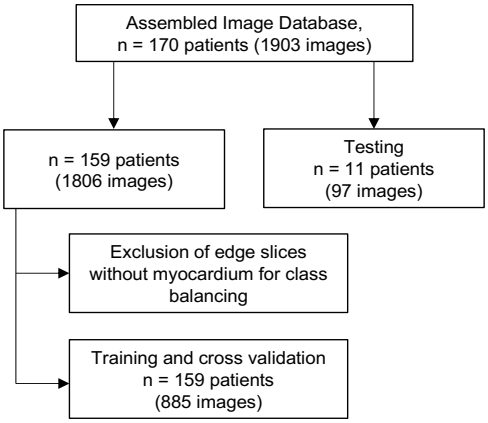

Figure 1: Flow chart showing study design for image assembly and distribution to training, cross-validation and testing.

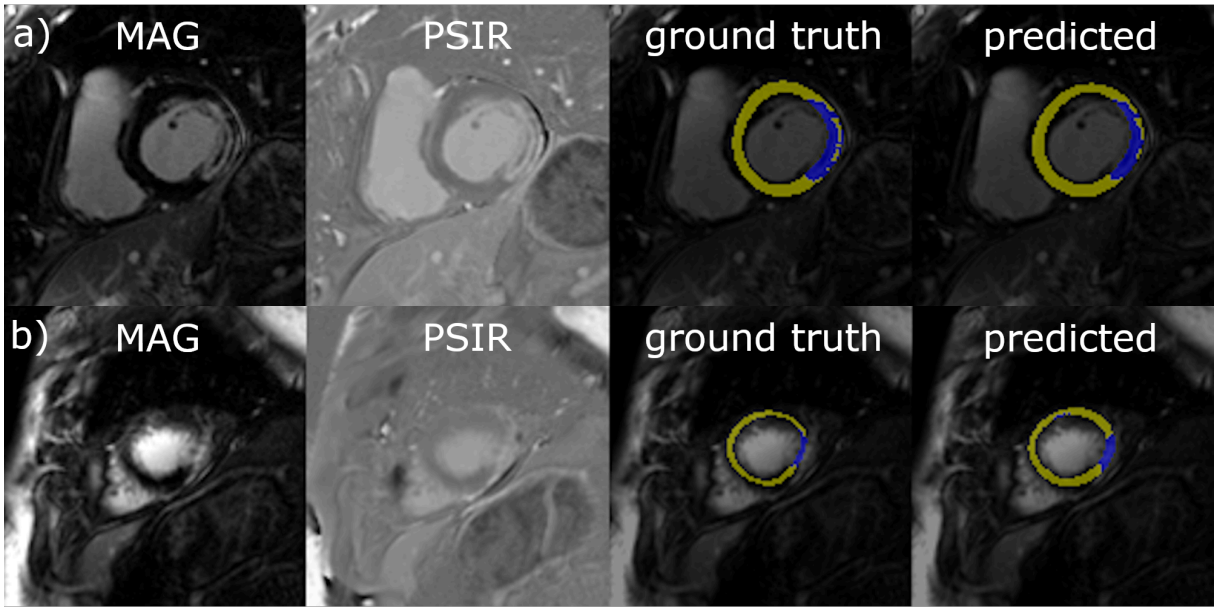

Figure 2: Exemplary illustration of automatic segmentation performed by nnU-net for the DUAL group. In a) an example with good performance (DSC_{myo} = 0.89, DSC_{scar} = 0.89) is depicted. No reflow areas were added to infarct zone as labelled in the training data. The second example (b) shows a specimen with lower performance of the trained model as measured by the DSC (DSC_{myo} = 0.67, DSC_{scar} = 0.33). Myocardium was interpreted thicker as in the manual reference, which both affected the results form myocardium and infarct zone. MAG, magnitude reconstruction; PSIR, phase sensitive inversion recovery.

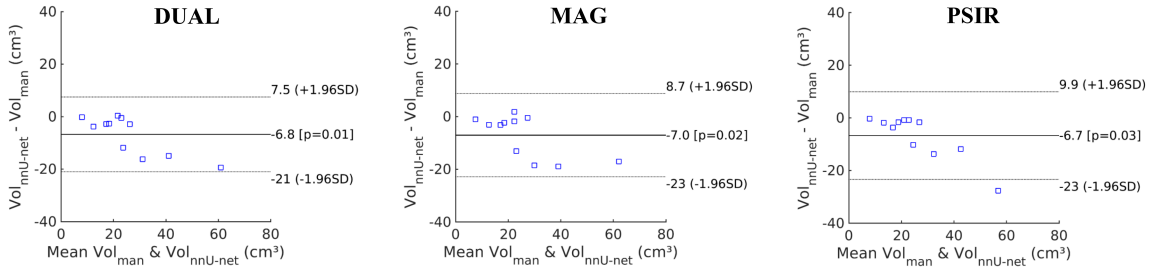

Figure 3: Bland Altman Analysis of Infarct Volumes. A general underestimation of the trained model is apparent. Underestimation was statistically significant compared to the manual ground truth (indicated by p-value). SD, standard deviation; Vol, volume; man, manual; MAG, magnitude reconstruction; PSIR, phase sensitive inversion recovery; DUAL, dual input of MAG and PSIR.

Tables

Table 1: Summary of patient demographics

		overall	training	testing
n patients		78	68	10
men		64	58	5
	mean age	64.3		
	age range	35 – 87		
women		14	8	5
	mean age	64.1		
	age range	48 - 77		
n examinations		170	159	11
	acute MI (0-4 days)	61	56	5
	subacute MI (7 days)	59	58	1
	chronic MI (12 months)	50	45	5
n images		1902	1806	96
	presence of infarction		699 (79 %)	72 (75 %)

Data of patients and examinations is shown in absolute numbers and regarding patient age as mean. Patient age is shown in years.

Table 2: Comparison of trained nnU-net predictions for myocardium and infarct zone using PSIR, magnitude or both contrasts as input.

	myocardium per slice	myocardium per patient	infarct zone per slice	infarct zone per patient
PSIR	0.80 ± 0.11	0.82 ± 0.03	0.66 ± 0.22	0.71 ± 0.08
MAG	0.81 ± 0.09	0.82 ± 0.03	0.67 ± 0.23	0.72 ± 0.10
DUAL	0.81 ± 0.10	0.83 ± 0.03	0.67 ± 0.24	0.72 ± 0.08

Data shows Dice-similarity coefficients as mean ± standard deviation across all patients and all slices, respectively. Results are shown for the two classes myocardium and infarct zone, calculated for the nnU-nets trained by the different groups PSIR, MAG and DUAL. PSIR, phase sensitive inversion recovery; MAG, magnitude; DUAL, input of PSIR and MAG.