MELBA Journal Machine Learning for Biomedical Imaging

The MYOSAIQ Challenge: Myocardial Segmentation with Automated Infarct Quantification

Olivier **Bernard** [1,2], William A. Romero R. [1], Cyprien Bouton [1], Celia Goujat [1], Hang Jung Ling [1], Pierre-Marc Jodoin [3], Fumin Guo [4], Calder Sheagren [5], Graham Wright [5], Abdul Qayyum [6], Moona Mazher [7], Steven A. Niederer [6], Hairui Wang [7], Xiaomei Wu [8], Franz Thaler [9,10], Gernot Plank [9], Martin Urschler [11], Ricardo M. Rosales [12], Esther Pueyo [12], Nicolas Duchateau [1,2], Frederic Cervenansky [1], Patrick Clarysse, Loic Belle [13], Thomas Bochaton [14,15], Nathan Mewton [14,15], Magalie Viallon [1,16], Pierre Croisille [1,16]

1 INSA-Lyon, Universite Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, F-42023, Saint Etienne, France
2 Institut Universitaire de France (IUF), France
3 Department of Computer Science, University of Sherbrooke, Sherbrooke, QC, Canada
4 Wuhan National Library of Optoelectronics, Huazhong University of Science and Technology, Wuhan, China
5 Sunnybrook Research Institute, University of Toronto, Toronto, Canada
6 National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, United Kingdom
7 Centre for Medical Image Computing, Department of Computer Science, University College London, London, United Kingdom
8 Fudan University, Shanghai, 200433, China
9 Gottfried Schatz Research Center: Medical Physics and Biophysics, Medical University of Graz, Graz, Austria
10 Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria
11 Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria
12 Instituto de Investigación en Ingeniería de Aragón, Universidad de Zaragoza; Instituto de Investigación Sanitaria de Aragón (IIS Aragón); CIBER-BBN, Zaragoza, Aragón, Spain
13 Centre Hospitalier Annecy Genevois, France
14 Hôpital Universitaire Cardiologique Louis Pradel, Bron, France
15 CarMeN INSERM U1060, Universite Claude Bernard Lyon 1, Lyon, France
16 Hôpital Universitaire de Saint-Etienne, Saint-Etienne, France.

Abstract

Late gadolinium enhancement (LGE) cardiac magnetic resonance (MR) imaging is the modality of choice to assess myocardial infarction (MI) lesions. Nowadays MI volume quantification is not performed routinely in clinical practice. Numerous deep learning (DL) methods have been developed to automate the segmentation of the myocardium and infarct regions. However, most studies rely on relatively small datasets which typically undergo pre-processing steps to standardize images and focus on a specific phase of myocardial infarction following reperfusion therapy. These limitations have impeded the development of models that are generalizable across diverse conditions and thus suitable for routine clinical use. To advance research and establish benchmarks in generalizable learning for myocardial infarct quantification, this paper presents findings from the Myocardial Segmentation with Automated Infarct Quantification (MYOSAIQ) challenge. The dataset set up for the challenge combines 439 CMR volumes from two multicenter clinical trials, with representative data acquired in acute and chronic phases after acute MI. Data were acquired in 16 centers using MRI scanners from three different vendors. Six teams participated until the end of the challenge, employing various baseline models, data augmentation techniques, and confidence strategies. To enhance the significance of this study, we compare the challengers' results with those of fine-tuned foundation models. Our results indicate that well-designed UNet-based techniques outperform fully automatic foundation models for LGE MR segmentation. While the best methods achieve high-quality and stable delineations of the left ventricle and myocardium under various conditions, they remain improvable in accurately segmenting infarct regions.

Keywords

Cardiac imaging, Late Gadolinium Enhancement, segmentation, myocardial infarct

Article informations

https://doi.org/10.59275/j.melba.2026-e93b
Volume 2026, Received: 2025-12-122, Published 2026-08-28
Corresponding author: olivier.bernard@insa-lyon.fr

©2026 Bernard. License: CC-BY 4.0
Check for updates

1. Introduction

Late Gadolinium Enhanced Cardiac Magnetic Resonance Imaging is a critical tool in clinical practice for evaluating myocardial damage following myocardial infarction (MI). This imaging technique is the only one enabling the quantification of myocardial tissue with abnormal gadolinium contrast washout, thereby highlighting areas of damaged myocardium. Accurately identifying these hyper-enhanced regions allows for precise determination of lesion size post-MI, a key factor for diagnosis, prognosis, and guiding therapeutic decisions (Ibanez et al., 2019). Moreover, the evolution of the size and features of the damaged tissue during the various inflammatory phases at the acute stage (within 8 days post-reperfusion) to the final infarct scar at the chronic stage (at least 1 month after percutaneous coronary intervention and tissue reperfusion) are key informations for patient recovery and follow-up characterization (Yellon et al., 1983; Ibanez et al., 2019). Thus, it is necessary to reliably estimate the lesion and infarct size at various stages over time. Finally, during the acute phase, blood flow may not be adequately restored into some regions of the myocardium despite successful reperfusion, resulting in post-reperfusion no-reflow lesions also called microvascular obstruction (MVO). Measuring the size of these no-reflow lesions is also crucial for understanding the progression of myocardial injury over time (Ibanez et al., 2019), but it also poses new challenges for segmentation (Bulluck et al., 2016).

While quantification is clinically valuable and necessary, it is rarely performed in routine practice. This is mainly due to the difficulty of accurately segmenting MVO and infarct regions on LGE images. Indeed, contrast kinetics and imaging physics lead to blurred, time-dependent boundaries that are highly susceptible to artifacts, particularly at the infarct core–border zone interface. Moreover, the small size of these structures relative to the current image resolution further amplifies relative segmentation errors. In practice, the annotation process typically requires substantial manual input from clinicians to ensure accurate and consistent delineation of myocardial and lesion boundaries across all image slices. As a result, automating this tedious and time-consuming task has become an active area of research, driven by recent advances in magnetic resonance (MR) image segmentation (Chen et al., 2020). Indeed, the advent of the deep learning paradigm has motivated the development of several neural network-based techniques to improve late gadolinium enhancement (LGE) cardiac MR segmentation (Lalande et al., 2022). However, most of these techniques have been trained and evaluated on relatively small datasets ($\leq$150 patients), with cardiac imaging samples collected from a single clinical center. These studies typically rely on a unique 2D multi-slice, multi-breathhold LGE acquisition strategy, combined with pre-processing steps to standardize images, correct for motion artifacts between slices, and focus on a specific phase of myocardial infarction. While these studies serve as proof-of-concept and demonstrate the potential and feasibility of extracting meaningful MI features from LGE images under constrained conditions, none has yet been successfully translated into routine clinical practice or scaled to larger, more realistic datasets. In particular, they have not been validated on multicenter, multi-vendor cohorts comprising both acute and chronic cases, acquired using a broader range of LGE techniques and imaging protocols. It is therefore crucial to establish frameworks (i.e. databases and evaluation platform) that enable the study and refinement of the generalization capacity of the best segmentation methods.

Open-access datasets strongly drive research progress on specific medical challenges. Earlier challenges (DE-MRI and MyoPS) on LGE segmentation have been conducted in small datasets ($\leq$45 patients) with images acquired under the same conditions (Karim et al., 2016a; Li et al., 2023). The recent LGE CMR dataset from the Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI (EMIDEC) (Lalande et al., 2022) stands as the most complete challenge and has been widely used to build and test the most advanced implementations of deep neural networks for the segmentation of cardiac structures in LGE CMRI. The top performing techniques (Lalande et al., 2022) rely on a cascaded two-stage framework based on convolutional neural networks (CNN), where the left ventricle and the myocardium are first segmented, followed by the delineation of the MI lesion and the MVO (Lecesne et al., 2023). The obtained results show high segmentation scores for the left ventricle and the myocardium with encouraging but improvable scores for MI lesions. However, the EMIDEC dataset collected in a single center contains only 100 patients with unstaged lesions, and using the same LGE imaging technique. A pre-processing procedure was also applied to compensate for the motion artifacts between non-contiguous 2D slices.

In this context, the MYOcardial Segmentation with Automated Infarct Quantification (MYOSAIQ) challenge was proposed and organized as part of the Functional Imaging and Modeling of the Heart (FIMH) conference in 2023 (endorsed by the MICCAI society), and was extended to a second phase in 2024. This challenge was set up to evaluate and compare the performance of automatic DL methods for quantifying myocardial LGE lesions observed after ischemia and reperfusion, at different phases of the longitudinal evolution of the disease. MR data were acquired during the acute phase and at delayed or chronic phases. Acute phase refers to 4-8 days post-MI and reperfusion therapy, where gadolinium contrast distributes over the reperfusion lesion, highlighting an area mixing oedema, apoptotic or already

necrotic tissue while also revealing MVO lesions as hypo-enhanced areas. Delayed or chronic phases refer to 1 month and 12 months post-MI and reperfusion, where the gadolinium contrast aggregates in the resulting final necrosis. Each data was annotated from a consistent contouring standard operating procedure described in section 2.2.

In this context, the main contributions of this collaborative paper are:

1. We release a new fully annotated dataset comprising 439 LGE MR images specifically designed for myocardial infarct quantification. The value of this dataset lies in its heterogeneity including the main LGE variants imaging techniques, acquired in 16 centers with 3 MR vendors, without any pre-processing steps, and obtained either in the acute phase or at later chronic time points.
2. We present the results from six distinct teams that participated in the MYOSAIQ challenge, each employing various baseline models, data augmentation techniques, and confidence strategies.
3. We compare the challengers' results with those of foundation models for LGE MR images based on the well-known SAM architecture (Kirillov et al., 2023). These foundation models were specifically designed to address the generalization problem, yet their performance has not been comprehensively compared to state-of-the-art methods for myocardial infarction quantification. While these models are semi-automatic and require initialization—which may introduce bias into the analysis—we found it worthwhile to include this comparison to better contextualize the potential of such approaches.
4. We conduct a comprehensive analysis considering multiple aspects, including overall segmentation quality, the impact of time following reperfusion therapy, the comparison between CNN and foundation models, and the effectiveness of confidence strategies.

2. Challenge framework

2.1 The MYOSAIQ data collection

The complete dataset includes 439 LGE examinations from two distinct multicenter clinical cohorts, designed to quantify MI lesions across different phases of disease progression. These phases encompass assessments conducted 4–8 days [D8] post-MI at the acute phase (MIMI cohort) and at 1 month [M1] and 12 months [M12] post-MI and reperfusion (HIBISCUS cohort). Consequently, the dataset is divided into three subgroups corresponding to these phases:

- D8: 123 patients (98 for training, 25 for testing, using a random split) with LGE images depicting acute myocardial infarction within 8 days post-MI (MIMI cohort). These patients underwent treatment via two different procedures: i) Percutaneous Coronary Intervention (PCI) with immediate stenting, or ii) the Minimalist Immediate Mechanical Intervention approach Belle et al. (2016).
- M1: 187 (155 for training, 32 for testing, using a random split) patients with LGE images acquired at 1 month post-PCI and reperfusion (HIBISCUS-STEMI cohort).
- M12: 129 (105 for training, 24 for testing, using a random split) patients with LGE images captured at 12 months post-PCI and reperfusion (HIBISCUS-STEMI cohort).

2.1.1 The MIMI cohort

The MIMI cohort (ClinicalTrials.gov Identifier: NCT01360242) is derived from a multicenter randomized trial involving over 16 centers, designed to compare immediate stenting with 24–48-hour delayed stenting in patients undergoing percutaneous coronary intervention (Belle et al., 2016). The MR protocol included T1, T2, T2*, Cine, and 3D/2D LGE scans. The MIMI cohort consists of patients admitted to French hospitals with ST-segment elevation myocardial infarction (STEMI) within less than 12 hours of symptom onset. These patients were randomized into two groups: immediate stenting ($n = 65$) or delayed stenting ($n = 58$) following the restoration of Thrombolysis In Myocardial Infarction (TIMI) grade 3 flow via thrombus aspiration. Patients in the delayed stenting group underwent a second coronary angiography for stent implantation, with a median delay of 36 hours (interquartile range: 29–46 hours) after randomization. The primary endpoint evaluated was MVO, expressed as a percentage of left ventricular myocardial mass, assessed using cardiac magnetic resonance imaging performed 5 days after the initial procedure (interquartile range: 4-6 days). Although a trend towards lower microvascular obstruction was observed in the immediate stenting group compared to the delayed stenting group (1.88% versus 3.96%; $P = 0.051$), this trend became significant after adjusting for the area at risk ($P = 0.049$) (Belle et al., 2016).

2.1.2 The HIBISCUS-STEMI cohort

The HIBISCUS-STEMI cohort ClinicalTrials.gov Identifier: NCT03070496) consists of consecutive patients admitted to the University Hospital of Lyon (HCL), a tertiary referral center, with suspected acute STEMI between 2016 and 2019. A total of 300 patients were included, all of them provided written informed consent. STEMI was defined according to the European Society of Cardiology guidelines, based on the presence of clinical symptoms (chest pain) associated with ST elevation > 2 mm in two contiguous leads on a standard 12-lead electrocardiogram, along with a significant elevation in troponin-I levels. All patients un-

derwent coronary angiography upon admission, followed by reperfusion via primary percutaneous coronary intervention. A complete myocardial enzyme release assessment was performed for all patients, who also underwent LGE cardiac MR imaging 1 month and 12 months after the acute myocardial infarction. All individual clinical, treatment, and outcome data were prospectively stored. Adverse clinical events were recorded during follow-up visits scheduled at 1 month, 1 year, and 2 years after the hospitalization. The MIMI and HIBISCUS-STEMI cohort were both approved by institutions's Review Board and Ethics Committee.

2.1.3 LGE imaging protocols

For LGE imaging in cardiac MR, the imaging protocols typically rely on inversion-recovery (IR) prepared T1-weighted sequences, usually performed 10 minutes after the intravenous administration of gadolinium chelate. When the sequence settings are optimized, a five-fold difference in intensity can be achieved between viable and nonviable myocardium (Amado et al., 2004). The recommended and most commonly used sequences were established as follows: (i) two-dimensional (2D) IR-prepared segmented gradient-echo (GRE) with magnitude and phase-sensitive inversion recovery (PSIR) reconstruction, and (ii) the three-dimensional (3D) IR-prepared segmented GRE with magnitude reconstruction (Viallon et al., 2011a). The 2D-PSIR sequence is acquired over multiple breath-holds, which may be inconsistent in some patients (e.g., due to varying lung volume), whereas the 3D-IR-GRE sequence is acquired over a single, but longer, breathold. When feasible in patients, the latter also results in a higher contrast-to-noise ratio (CNR) between the gadolinium enhanced lesion and the normal myocardium, being the most performant sequence for LGE quantification (Viallon et al., 2011b). The MYOSAIQ dataset included 13% and 87% of 2D-PSIR and 3D-IR-GRE sequences, respectively. The sequence type information is not disclosed to participants to encourage the generalization of the deployed methods. All LGE cardiac MR images were acquired in DICOM format using 1.5 Tesla whole-body clinical multivendors MRI scanners. An in-house software was then used to transform the images into the NIfTI format (3D images), representing the final files delivered to the challenge participants.

2.2 Data annotation procedure

Every LGE cardiac MR image was manually annotated using the 3D Slicer software. An expert with over 20 years of experience in cardiac radiology and MRI performed the annotations for the entire dataset. To enable the evaluation of inter-observer variability, a second expert with 10 years of experience independently annotated the test set. The semi-automatic Full Width at Half Maximum (FWHM) approach

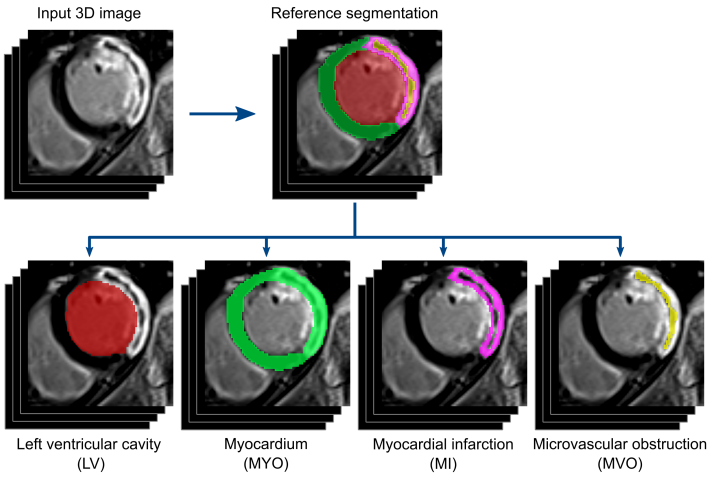


Figure 1: Reference segmentation and regions of interest: left ventricular cavity (red), myocardium (green), myocardial infarction (magenta) and microvascular obstruction (yellow).

currently recommended in the literature guided the experts but results were always systematically supervised and corrected if needed during the expert annotation process. Four main regions were considered: the left ventricle (LV) cavity, the left ventricle myocardium (MYO), the myocardial infarction (MI) and the microvascular obstruction (MVO) as presented in Figure 1. As such, the MYO corresponds to the entire myocardium, including both the healthy tissue and the regions affected by both MI and MVO when present. Particular attention was given to the delineation of each structure, especially at the base and apex, where the contours are prone to variations among experts (Bernard et al., 2018). To generate consistent annotations for the research community, we chose to apply the standard operating procedures that were already used by the EMIDEC challenge, which rely on the following constraints:

- The LV cavity must be completely covered, including the papillary muscles,
- MI refers to tissue death (infarction) of the heart muscle, and must be located inside the myocardium,
- MVO is characterized as hypo-enhanced regions, appearing as black areas surrounded by a bright rim within the myocardium. During the contouring of this structure, black signals caused by noise or artifacts should be excluded.

The reference-labeled segmentations were stored in the NIfTI format with the following labels: background $= 0$, LV cavity $= 1$, healthy myocardium $= 2$, MI $= 3$, and MVO $= 4$.

		Territory	D8	M1	M12
Clinical metrics					
Transmurality (Extent of a lesion across the thickness of the muscle)	(%)	LAD	40 ± 16	27 ± 15	25 ± 16
		LCX	29 ± 09	17 ± 09	12 ± 08
		RCA	24 ± 10	17 ± 09	12 ± 08
Endocardial Surface Length	(%)	LAD	34 ± 11	29 ± 13	24 ± 14
		LCX	33 ± 08	22 ± 12	19 ± 11
		RCA	31 ± 09	18 ± 09	16 ± 08
Lesion Volume	(mL)	LAD	34 ± 12	18 ± 11	17 ± 12
		LCX	30 ± 10	13 ± 08	12 ± 08
		RCA	23 ± 14	11 ± 07	11 ± 06
Lesion Size	(%)	LAD	33 ± 10	19 ± 10	18 ± 09
		LCX	25 ± 07	14 ± 07	11 ± 06
		RCA	27 ± 08	14 ± 09	12 ± 07

Table 1: Descriptors of the myocardial infarction lesion categorized by territory and time post reperfusion therapy. The endocardial surface length and the lesion size are provided in proportion of the whole endocardial length and the myocardial area, respectively.

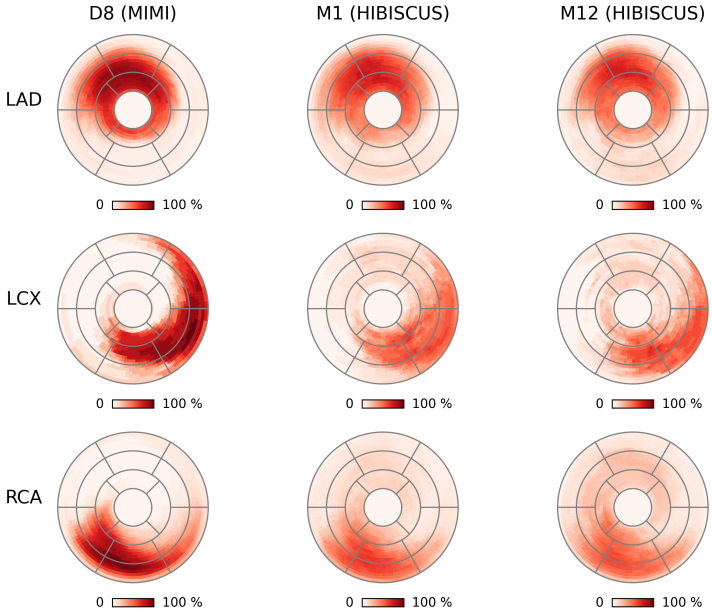


Figure 2: Infarct location of the LAD, LCX and RCA infarcts after 8 days, 1 month and 1 year post reperfusion (average across the subjects in each subgroup).

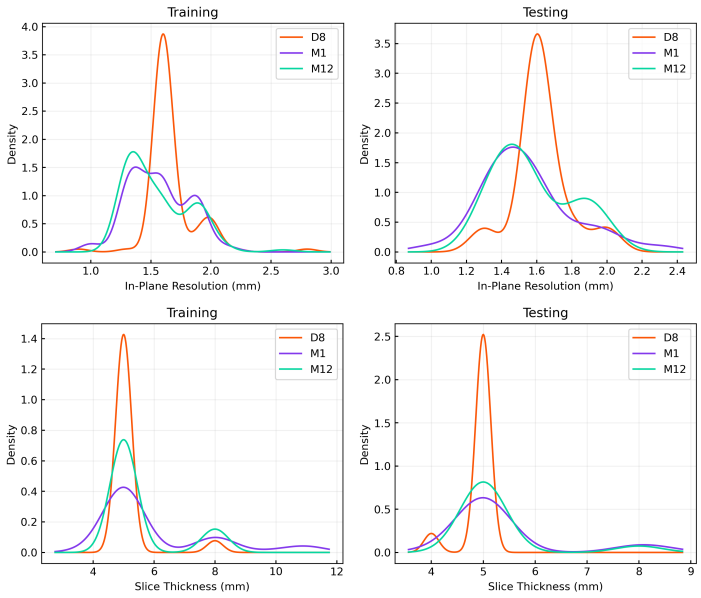


Figure 3: Distribution of the in-plane resolution and slice thickness for the training and test datasets across the three experimental configurations (D8, M1, and M12).

2.3 Image-based clinically relevant descriptors

To describe the properties of the MYOSAIQ dataset, Table 1 presents the key descriptors of myocardial infarct lesions derived from expert annotations. These descriptors are categorized based on the affected myocardial territory—left anterior descending (LAD), left circumflex artery (LCX), and right coronary artery (RCA)—as well as the time points following reperfusion therapy (day 8 [D8], month 1 [M1], and month 12 [M12]). The extent of myocardial damage is expressed as a proportion of the total myocardium or the total endocardial surface area. In addition, Figure 2 provides an overview of the mean infarct location for each affected myocardial territory and time point. For each subgroup, the corresponding Bull's Eye display was computed using the method described in (Duchateau et al., 2023).

Finally, Figure 3 illustrates the distributions of the in-plane resolution and slice thickness for the three experimental configurations (D8, M1, and M12), separately for the training and test datasets. As shown in the figure, the spatial resolution distributions remain highly comparable across all configurations. The in-plane resolution is consistently distributed between approximately 1.3 and 1.9 mm, while the slice thickness is centered around 5 mm in both the training and test datasets. These observations indicate that the spatial sampling characteristics are well matched between the training and test datasets, thereby limiting potential differences related to voxel spacing.

2.4 Evaluation platform

A Codalab environment (Pavao et al., 2023) was deployed as the evaluation platform for the segmentation challenge, offering a robust and transparent system for managing and assessing participants' submissions. The leaderboard allowed participants to monitor their progress and benchmark their results against others in real-time during both the training and testing phases. During the training phase (4 months), participants could evaluate their segmentations us-

ing the training dataset and corresponding reference masks, providing an opportunity for learning and experimenting with the platform. No restriction was imposed on the number of submissions during this phase, enabling extensive refinement of methods. In contrast, the testing phase (1 month) limited participants to a maximum of five submissions on the testing dataset, with no access to the reference masks, ensuring a fair and competitive environment. The scoring code used on the online platform was released to all participants two months prior to the competition's launch. The source code, developed in Python, was published in a public repository[1] to ensure transparency and provide participants with sufficient time to familiarize themselves with the evaluation criteria and metrics. This platform is now accessible and will remain active and maintained as long as the data remains relevant for clinical research. A complementary set of metrics was employed to evaluate the quality of the automatically segmented masks in comparison to the ground truth, as detailed below:

2.4.1 Geometrical metrics

The standard Dice, the maximum Hausdorff surface distance (HD) and Average Symmetric Surface Distance (ASSD) were used to compute geometrical scores for the four labels, i.e. the LV cavity, the full myocardium, the MI and the MVO. All these metrics were computed from the 3D masks.

2.4.2 Quality control metrics

We also implemented four indices to assess the quality of the estimation of the myocardial infarction lesion metrics, namely the correlation (CC), the mean absolute distance (MAE), the limits of agreement (LOA) and the continuous ranked probability score (CRPS) inspired by a Kaggle challenge Gneiting and Raftery (2007). While the first three metrics assess the accuracy of the extracted volumes, the CRPS evaluates the quality of the models' confidence in their predictions. In practice, each participant was asked to calculate the volumes and the cumulative probability distribution for each of the four cardiac structures. The CRPS was then computed for each structure as follows:

$$CRPS = \frac{1}{600N} \sum_{m=1}^{N} \sum_{n=0}^{599} \left(P(y \leq n) - \mathbb{1}_{\{n \geq V_m\}} \right)^2, \quad (1)$$

where P is the predicted cumulative distribution, n is a value ranging from 0 to 599 (we assume that the estimated volumes do not exceed 599 mL, which is consistent with human physiology, even in the presence of pathologies), N denotes the number of patients in the dataset under evaluation, V_m is the reference volume (in mL) for patient m and $\mathbb{1}_{\{n \geq V_m\}}$ is the indicator function, which equals 1 if $n \geq V_m$, and 0 otherwise. Thus, if a participant was able to estimate a volume along with its confidence value for a given structure, it is straightforward to compute the associated cumulative distribution under the Gaussian assumption, where the volume and confidence values are treated as the mean and standard deviation of the distribution. However, participants may provide distributions that do not follow a Gaussian pattern, based on their own strategy. If a participant did not provide any CRPS value, the cumulative distributions associated with each estimated volume were automatically calculated and corresponded to Heaviside step functions centered on the volumes computed automatically from the segmentation estimates.

Finally, if a participant's MVO prediction was missing for a given patient, no value was assigned to the corresponding geometric and clinical metrics. Consequently, these cases were excluded from the average values reported in the results section table. However, to ensure fairness, we also provide the number of missed cases for each method to assess their sensitivity.

2.4.3 Leaderboard

A leaderboard was computed during the first phase of the challenge. Two separate overall rankings were established based on (i) geometric metrics (Dice, ASSD, HD) and (ii) clinical metrics (CC, MAE, LOA, CRPS). For each ranking, a score was computed by summing the ranks obtained by each method across all metrics and cardiac structures (LV, MYO, MI), where ranks ranged from 1 (best) to N (worst), with N denoting the number of participants. MVO was not included in the ranking computation, as the methods do not segment the same number of patients. A dedicated analysis of MVO segmentation performance was, however, conducted and discussed separately during the FIMH 2023.

2.5 Backbones architectures

All teams employed UNet-inspired backbone architectures, including nnUNet (P1), 3D and 2D UNet variants (P2 and P6), UNet++ (P5), xLSTM-UNet (P4), and MedNeXt (P3). This diversity resulted in a broad spectrum of model complexities, ranging from 1.4 million parameters (P4) to 50.3 million parameters (P5) per base model. xLSTM-UNet features a UNet-based structure enhanced with a parallel cross-window module to capture long-range dependencies (Beck et al., 2024), along with a cross-attention transformer block to compute novel feature representations using cross-attention mechanisms. Team P3 implemented a model based on the MedNeXt architecture (Roy et al., 2023). To improve performance, each convolutional block was modified with separable convolutional kernels and inverted bottleneck layers, drawing inspiration from the design principles

1. https://github.com/creatis-myriad/myosaiq-challenge

Table 2: Summary of the main characteristics of the submitted techniques, including backbone architectures, data augmentation strategies, and confidence estimation approaches. Abbreviations: Rotations (R), Flipping (F), Scaling (S), Deformations (D), Gaussian Noise (GN), Brightness (B), Gamma correction (G). Methods shown in blue correspond to publicly available implementations and provide direct hyperlinks to the associated code repositories.

Method	Backbone architecture	Number of parameters	Data augmentation							Synthesis	Confidence strategy
			Spatial augmentations				Intensity augmentations				
			R	F	S	D	GN	B	G		
P1	nnUNet	16.1 M	✓	✓	✓		✓	✓	✓	3D GAN	Logit calibration
P2	Cascaded 3D UNet	2.6 M	✓		✓	✓		✓		No	Ensemble with 5 models
P3	MedNeXt	5.6 M	✓	✓	✓	✓	✓	✓	✓	No	Ensemble with 5 models
P4	xLSTM-UNet	1.4 M	✓	✓	✓	✓	✓	✓	✓	No	No
P5	Cascaded UNet++	50.3 M	✓		✓					No	Ensemble with 8 models
P6	Cascaded 2D nnUNet	7.4 M	✓	✓	✓		✓		✓	No	Ensemble with 10 models

of ConvNeXt (Liu et al., 2022). These modifications aim to maintain rich semantic information across the network. Team P5 used the UNet++ (Zhou et al., 2020) architecture with eight different backbones based on the widely available VGG, ResNet, and DenseNet methods. The eight segmentation predictions were fused for each class using a STAPLE algorithm (Warfield et al., 2004), and the resulting probability maps were entered into a multi-class continuous min-cut algorithm to generate the final segmentation with five classes dimensions.

Three teams (P2, P5, P6) employed a cascaded strategy where the first stage was used to detect the LV and the full myocardium. This result was then concatenated with the original image to create the input data given to the second stage which focused on the segmentation and the separation between healthy and damaged myocardial tissue. In the case of D8 dataset, a third stage was employed to separate between MI and MVO. Teams P2 and P5 trained their cascaded models in an end-to-end manner, whereas P6 trained each base UNet separately. Note that for these methods, the total number of parameters corresponds to the values presented in Table 2, multiplied by the number of levels involved in the cascade strategy.

2.6 Data augmentation

All participants in the challenge employed various data augmentation techniques to enhance their models. Two main categories of data augmentations were considered: (1) spatial transformations, which increase the sample size by applying rotations, flips, scaling, or deformations to the original images; and (2) intensity-based techniques, which preserve the spatial configuration of anatomical structures while altering their visual appearance. Team P1 employed a 3D GAN-based data augmentation strategy to train their nnUNet model using a combination of original and synthetic data. In particular, a 3D-UNet generator and a PatchGAN discriminator were employed to train the GAN model. Each sample was preprocessed through resampling, padded cropping, and Z-score normalization. For synthetic data generation, the training labels were deformed using both rigid and non-rigid transformations.

2.7 Confidence strategy

Five teams developed strategies to enhance the confidence in their clinical metric estimates. Team P1 calibrated their nnU-Net model output using a temperature scaling procedure (Guo et al., 2017) to generate probabilities that better reflect true values. These probabilities were then used to compute a volume-wise probability function and its cumulative density function, which served to calculate the CRPS metrics 1. Teams P2, P3, P5, and P6 employed different ensemble strategies. Teams P2 and P3 trained five models and averaged their outputs to produce the final class predictions for each image. Similarly, team P5 trained eight models but replaced the averaging step with a merging procedure based on the STAPLE algorithm (Warfield et al., 2004). Team P6 trained ten models to compute the mean and standard deviation maps for each class and image. These maps were then used to define Gaussian distributions of the estimated volumes, which were employed to calculate the CRPS metrics.

2.8 Foundation models

The application of foundation models for medical image segmentation has seen a rapid expansion in recent years, driven by the advent of the SAM model developed for natural image processing (Kirillov et al., 2023). These methods are semi-automatic and require initialization using prompts such as bounding boxes, points, masks, or text. MedSAM (Ma et al., 2024) is one of the most advanced model for medical image segmentation. It was developed on a large-scale medical image dataset with 1.6 M image-mask pairs, covering 10 imaging modalities. This large-scale dataset allows MedSAM to learn a rich representation of medical images, capturing a broad spectrum of anatomies and lesions

across different modalities. The dataset has been created by collating images from publicly available medical image segmentation datasets, which were obtained from various sources across the internet. MedSAM was built upon the SAM model by fine-tuning the bounding box prompting and the decoder, resulting in a 91M parameter model. In the context of LGE-MRI, we recently introduced mcMedSAM, a model designed to segment multiple cardiac structures from a single bounding box centered on the left ventricle (Goujat et al., 2025). This model extends the original MedSAM framework by incorporating structure-specific tokens and a hierarchical end-to-end architecture specifically tailored for multi-class segmentation. We also demonstrated that achieving competitive performance relative to state-of-the-art models such as nnU-Net requires the bounding box to be positioned with high spatial accuracy (within 3 mm in the x and y axes) (Goujat et al., 2025). In this study, we removed the bounding-box prompt entirely to evaluate mcMedSAM in a fully automated setting. We froze the encoder parameters inherited from the MedSAM architecture and fine-tuned the decoder on the MYOSAIQ training dataset without data augmentation, enabling a fair comparison with other automatic segmentation methods. We refer to this configuration as mcMedSAM-auto. For completeness, and to support the discussion, we also report results for the original MedSAM model, whose decoder was fine-tuned on the training dataset without data augmentation to better adapt to LGE images. This semi-automatic version, referred to as MedSAM-f-semi, was initialized—during both training and inference—with bounding boxes generated from the reference masks and randomly perturbed along the horizontal and vertical axes by uniform shifts of up to ± 1 mm on each side.

3. Results

The quality of the results obtained by each method was compared with the inter-observer variability on the testing dataset (denoted as Inter-obs in Tables 3 and 5).

3.1 MYOSAIQ challenge results

Table 3 and 5 display the geometric and clinical scores obtained by all the participants (P1 to P6) and the foundation models (mcMedSAM-auto to mcMedSAM-semi) on the test dataset of the MYOSAIQ challenge. From these tables, we observe that the different participants' methods produced very similar scores, whatever the backbone architecture and the data augmentation strategy. This is further supported by the statistical analysis of the geometrical metrics, which confirms comparable performance levels. Paired t-tests were performed after verifying the normality of the paired differences. These findings are further supported by the bootstrap analysis reported in Table 4, where the narrow 95% confidence intervals indicate that the reported performance estimates are stable with respect to variations in the test cohort. Interestingly, segmentation performance decreases across the LV, MYO, and MI structures, reflecting increasing task complexity. This is evidenced by the average Dice scores of 0.927, 0.832, and 0.668, respectively, and corresponding HD values of 7.3, 9.6, and 19.5 mm. This trend is also observed in the clinical scores, with average correlation values of 0.977, 0.937, and 0.890. Regarding MVO, the results indicate that these structures are generally detected when present in the image and rarely misidentified when absent, demonstrating a well-balanced performance across methods. However, due to the diffuse nature of MVO, its segmentation remains challenging, as reflected by the average Dice score of 0.606 and HD score of 16.2 mm. Figures 4 and 5 illustrate two representative examples of the evaluated methods. The first example corresponds to a typical case where all three methods (P2, P5, and MedSAM Auto) closely match the expert annotation. The second example illustrates a more challenging case, highlighting complementary failure modes. While P2 and P5 tend to confuse microvascular obstruction with infarction and miss the expert-annotated infarcted region, MedSAM Auto successfully differentiates both pathological structures but exhibits less accurate localization. It is worth noting that no specific preprocessing tailored to the characteristics of LGE images, nor explicit post-processing steps, were consistently implemented across participants. Therefore, no conclusion can be drawn from this challenge regarding the potential impact or necessity of such pre- or post-processing strategies.

3.2 Foundation models' results

Tables 3 and 5 report the performance of the two foundation models evaluated in this study. The results show that the fully automatic model achieves competitive performance for the LV, slightly lower performance for the MYO, and clearly inferior performance for the MI and MVO when compared with the CNN-based models. In contrast, the semi-automatic model obtains high scores for the different structures, with geometric accuracy metrics such as a HD score of 4.8 mm for the LV, 7.4 mm for the MYO, 8.4 mm for the MI and 5.2 mm for the MVO, which are the lowest among all the models tested. Although these results are influenced by the initialization procedure, they highlight the critical role of the bounding box in such models and quantify the performance improvements brought by its use.

3.3 Effect of time elapsed since reperfusion therapy

Figure 6 presents boxplots of Dice and HD scores for each participant based on the time elapsed since reperfusion

Table 3: Mean and standard deviations of the geometrical metrics (Dice, ASSD and HD) for the participants P1-P6 and the foundation models (mcMedSAM and MedSAM-f) on the MYOSAIQ challenge test set. **Bold** face numbers are the best results for the automatic methods in each column and blue numbers indicate results that are not statistically significantly different from the best result in each column (paired t-test, > 0.05). TP represents the number of times a MVO structure was present in an image and correctly detected. TN represents the number of times a method correctly identified that no MVO structure is present in an image. Given that the methods do not segment the same number of patients for MVO, we decided not to provide a ranking for this structure. Inter-obs corresponds to the inter-expert variability computed from two different experts on the same testing dataset.

Method	LV			MYO			MI			MVO				
	Dice	ASSD (mm)	HD (mm)	Dice	ASSD (mm)	HD (mm)	Dice	ASSD (mm)	HD (mm)	Dice	ASSD (mm)	HD (mm)	TP	TN
P1	0.928 ±0.031	0.5 ±0.2	7.2 ±2.4	0.840 ±0.036	**0.4** ±**0.2**	**9.0** ±**2.8**	0.669 ±0.177	**1.0** ±**1.7**	19.8 ±14.4	0.537 ±0.190	1.2 ±1.0	19.5 ±14.6	10/10	15/15
P2	**0.933** ±**0.029**	**0.4** ±**0.2**	6.8 ±2.1	0.840 ±0.038	**0.4** ±**0.1**	**9.0** ±**3.0**	**0.700** ±**0.145**	**1.0** ±**1.7**	**16.8** ±**13.2**	0.681 ±0.119	0.7 ±0.4	14.4 ±4.7	9/10	15/15
P3	0.930 ±0.034	0.5 ±0.4	7.7 ±5.8	**0.841** ±**0.041**	**0.4** ±**0.2**	9.9 ±6.1	0.688 ±0.168	**1.0** ±**1.3**	17.3 ±12.4	0.648 ±0.140	0.8 ±0.4	14.1 ±5.0	8/10	15/15
P4	0.926 ±0.031	0.5 ±0.2	7.3 ±2.4	0.827 ±0.043	**0.4** ±**0.1**	9.5 ±3.3	0.660 ±0.159	**1.0** ±**1.5**	19.5 ±13.5	0.647 ±0.126	0.7 ±0.3	13.5 ±5.0	8/10	15/15
P5	0.914 ±0.034	0.6 ±0.3	7.9 ±3.2	0.806 ±0.043	**0.4** ±**0.2**	10.6 ±3.4	0.628 ±0.170	1.2 ±1.8	23.9 ±14.9	0.552 ±0.186	1.0 ±0.5	16.2 ±3.4	9/10	15/15
P6	0.931 ±0.030	**0.4** ±**0.2**	**6.7** ±**2.2**	0.839 ±0.038	**0.4** ±**0.1**	9.6 ±3.8	0.661 ±0.150	**1.0** ±**1.3**	19.7 ±13.2	0.568 ±0.145	1.1 ±0.6	19.6 ±10.3	7/10	15/15
mcMedSAM auto	0.927 ±0.027	**0.4** ±**0.1**	7.7 ±1.9	0.794 ±0.041	**0.4** ±**0.1**	10.1 ±3.4	0.544 ±0.167	1.3 ±1.2	27.8 ±12.7	0.328 ±0.133	1.9 ±1.1	26.8 ±13.4	10/10	15/15
mcMedSAM semi	0.965 ±0.007	0.2 ±0.0	4.8 ±1.6	0.890 ±0.028	0.2 ±0.1	7.4 ±3.0	0.740 ±0.079	0.4 ±0.2	8.4 ±5.0	0.696 ±0.088	0.4 ±0.1	5.2 ±1.1	–	–
Inter-obs	0.954 ±0.041	0.3 ±0.3	5.7 ±4.5	0.902 ±0.083	0.2 ±0.2	7.5 ±5.2	0.906 ±0.100	0.2 ±0.3	9.2 ±10.7	0.937 ±0.088	0.1 ±0.1	2.0 ±2.1	–	–

therapy (day 8, month 1, and month 12). This figure shows that Dice scores remain stable over time for LV and MYO across all participants. A performance drop is observed for the MI region at D8 compared to M1 and M12, both in terms of median values and standard deviations. Similarly, HD scores for LV and MYO remain consistent across participants. However, MI results show greater variability: while teams P1, P2, and P3 maintain stable HD scores over time, teams P4, P5, and P6 exhibit a slight performance decline at M1 and M12. These results indicate the overall stability of the DL methods for the segmentation of LV and MYO structures over time and the increased complexity at D8 of the segmentation of the MI region, likely due to the presence of the MVO structures, which complicate infarct zone detection.

3.4 Confidence in the clinical scores

Confidence strategies can be evaluated through the analysis of the CRPS score (see Table 5). The ensemble-based methods (teams P2, P3, P5, and P6) produced similar CRPS scores for each structure, suggesting that the choice of strategy has little impact on confidence estimation in the clinical metrics. Interestingly, these results are comparable to those of team P4, for which no specific confidence estimation strategy was applied in their clinical metrics. This indicates that while ensemble strategies may improve segmentation accuracy, they are not particularly effective in capturing uncertainty. In contrast, team P1 employed a more advanced probabilistic framework based on a temperature-scaling procedure. Their approach proved highly effective, achieving CRPS scores 3 to 7 times lower than those of other participants. This demonstrates both the feasibility and the value of developing methods that are not only accurate but also capable of providing reliable confidence estimates.

4. Discussion and Conclusion

Table 4: Bootstrap analysis of the segmentation performance on the MYOSAIQ challenge test set. For each participant (P1–P6) and automatic version of the foundation model (mcMedSAM), the Dice, ASSD, and HD metrics are reported as *bootstrap mean*, *standard deviation*, and 95% *confidence interval* (CI), estimated from 5,000 bootstrap resamples of the test subjects. Confidence intervals are reported as $[CI_{\text{lower}}, CI_{\text{upper}}]$.

Method	LV			MYO			MI			MVO		
	Dice	ASSD (mm)	HD (mm)	Dice	ASSD (mm)	HD (mm)	Dice	ASSD (mm)	HD (mm)	Dice	ASSD (mm)	HD (mm)
P1	0.928; 0.004 [0.921, 0.935]	0.5; 0.0 [0.4, 0]	7.2; 0.3 [6.7, 7.7]	0.840; 0.004 [0.832, 0.848]	0.4; 0.0 [0.3, 0]	9.0; 0.3 [8.4, 9.6]	0.669; 0.02 [0.627, 0.706]	1.0; 0.2 [0.7, 0]	19.8; 1.6 [16.8, 23.2]	0.538; 0.063 [0.404, 0.652]	1.2; 0.3 [0.7, 0]	19.6; 4.6 [12.3, 29.5]
P2	0.933; 0.003 [0.926, 0.939]	0.4; 0.0 [0.4, 0]	6.8; 0.2 [6.4, 7.3]	0.840; 0.004 [0.831, 0.848]	0.4; 0.0 [0.3, 0]	9.0; 0.3 [8.4, 9.7]	0.700; 0.016 [0.668, 0.73]	1.0; 0.2 [0.7, 0]	16.9; 1.5 [14.1, 19.9]	0.681; 0.039 [0.601, 0.756]	0.7; 0.1 [0.5, 0]	14.4; 1.5 [11.6, 17.6]
P3	0.930; 0.004 [0.922, 0.937]	0.5; 0.0 [0.4, 0]	7.7; 0.7 [6.7, 9.2]	0.841; 0.005 [0.832, 0.85]	0.4; 0.0 [0.3, 0]	9.9; 0.7 [8.7, 11.4]	0.688; 0.019 [0.648, 0.723]	1.0; 0.2 [0.7, 0]	17.3; 1.4 [14.7, 20.2]	0.647; 0.05 [0.545, 0.741]	0.8; 0.2 [0.5, 0]	14.1; 1.8 [10.9, 17.8]
P4	0.926; 0.003 [0.919, 0.933]	0.5; 0.0 [0.4, 0]	7.3; 0.3 [6.8, 7.8]	0.827; 0.005 [0.818, 0.836]	0.4; 0.0 [0.4, 0]	9.5; 0.4 [8.8, 10.3]	0.659; 0.018 [0.623, 0.693]	1.0; 0.2 [0.7, 0]	19.5; 1.5 [16.6, 22.7]	0.647; 0.045 [0.555, 0.73]	0.7; 0.1 [0.5, 0]	13.5; 1.8 [9.9, 17.0]
P5	0.914; 0.004 [0.906, 0.921]	0.6; 0.0 [0.5, 0]	7.9; 0.4 [7.3, 8.6]	0.806; 0.005 [0.797, 0.816]	0.4; 0.0 [0.4, 0]	10.6; 0.4 [9.9, 11.4]	0.627; 0.019 [0.59, 0.663]	1.2; 0.2 [0.9, 0]	23.9; 1.7 [20.7, 27.2]	0.553; 0.062 [0.431, 0.674]	1.0; 0.2 [0.7, 0]	16.2; 1.1 [14.0, 18.3]
P6	0.931; 0.003 [0.924, 0.938]	0.4; 0.0 [0.4, 0]	6.7; 0.2 [6.2, 7.2]	0.839; 0.004 [0.83, 0.847]	0.4; 0.0 [0.3, 0]	9.6; 0.4 [8.8, 10.4]	0.661; 0.017 [0.626, 0.694]	1.0; 0.1 [0.7, 0]	19.7; 1.5 [16.9, 22.7]	0.568; 0.054 [0.462, 0.67]	1.1; 0.2 [0.7, 0]	19.7; 3.9 [12.9, 27.7]
mcMedSAM auto	0.927; 0.003 [0.921, 0.933]	0.4; 0.0 [0.4, 0]	7.7; 0.2 [7.2, 8.1]	0.794; 0.005 [0.784, 0.802]	0.4; 0.0 [0.4, 0]	10.1; 0.4 [9.4, 10.9]	0.543; 0.019 [0.505, 0.579]	1.3; 0.1 [1.1, 0]	27.8; 1.4 [25.1, 30.5]	0.328; 0.042 [0.247, 0.411]	1.9; 0.3 [1.3, 0]	26.8; 4.2 [19.1, 35.3]

4.1 How far are we from solving the LGE MRI segmentation problem ?

The results indicate that convolution-based approaches yield comparable geometric scores, which are often not statistically different, regardless of the backbone architecture or data augmentation strategy. Moreover, these methods performed similarly independent on when reperfusion therapy was performed, with some differences for the MI region. Although these observations require validation on larger cohorts, our results suggest that UNet-inspired backbone architectures may reach a performance plateau when relying solely on image data. While results obtained on the LV are competitive, it appears that the same level of accuracy is still difficult to obtain for the MYO, MI and MVO structures. It is therefore important to assess the performance of these methods relative to inter-observer variability, which motivated us to obtain manual annotations from a second expert for the 81 LGE images in the testing dataset. These additional annotations allow assessing inter-observer variability which further enhances the comprehensiveness of this study. The inter-observer variability measured on the test (Inter-obs) dataset is reported in the last two rows of Tables 3 and 5. Unsurprisingly, it appears that the MYO and MI are regions that are more difficult to annotate than the LV, with a HD score of 5.7, 7.5 and 9.2 mm, respectively. In terms of segmentation, the comparison between the participants' scores and inter-observer variability suggests that LV delineation has reached high levels of performance under controlled conditions, although challenges related to variability and generalization across clinical settings remain. However, varying levels of difficulty persist for MYO, MI, and MVO, with MI and MVO being the most challenging due to their lower contrast and higher variability. Regarding clinical indices, the results show that the best-performing methods are competitive with experts for estimating LV and MYO volumes, although improvements are still needed for MI and MVO. These results clearly highlight the need for further improvements in the quality of segmentation for the MI and MVO regions. In this regard, the computed inter-observer values can serve as benchmarks for the community to strive towards in future developments.

4.2 Where do foundation models stand ?

In this study, we compared the methods developed by the challenge participants with state-of-the-art foundation models. The fully automatic mcMedSAM model provided accurate LV segmentations but did not yield improvements for the other structures relative to CNN-based approaches. This outcome is likely due to the limited presence of LGE MR images—and thus the absence of MI, and MVO labels—in the pre-training datasets used for the foundation models. Moreover, the inherent flexibility of transformer-based architectures makes them less effective than CNNs when fine-tuned on relatively small datasets such as MYOSAIQ. In contrast, the high performance achieved by the semi-automatic fine-tuned model suggests that MedSAM's encoder captures generic and transferable features. This enables efficient adaptation of the decoder (2 million parameters), resulting in performance levels comparable to inter-observer variability across both geometric and clinical metrics. While these findings highlight the potential of foundation models to overcome some of the limitations of con-

Table 5: Clinical scores (CC, MAE, LOA, CRPS, see details in section 2.4.2) for the participants P1-P6 and the foundation models (mcMedSAM and MedSAM-f) on the MYOSAIQ challenge test set. **Bold** face numbers indicate the best results for the automatic methods in each column. Given that the methods do not segment the same number of patients for MVO, we decided not to provide a ranking for this structure. Inter-obs corresponds to the inter-expert variability computed from two different experts on the same testing dataset.

Method	LV				MYO				MI				MVO			
	CC	MAE (mL)	LOA (mL)	CRPS	CC	MAE (mL)	LOA (mL)	CRPS	CC	MAE (mL)	LOA (mL)	CRPS	CC	MAE (mL)	LOA (mL)	CRPS
P1	0.969	8.2	23.8	**0.002**	**0.967**	**8.6**	**24.5**	**0.004**	0.882	3.4	12.1	**0.001**	0.972	2.7	6.5	0.000
P2	**0.980**	**6.7**	**18.0**	0.011	0.946	11.2	31.1	0.018	0.892	3.4	12.3	0.005	0.989	1.1	3.0	0.003
P3	0.978	7.4	19.6	0.012	0.924	13.1	38.1	0.022	0.888	3.2	11.7	0.006	0.990	2.6	4.9	0.006
P4	0.977	7.5	20.4	0.012	0.918	12.3	38.2	0.021	0.868	3.3	12.6	0.006	0.997	1.4	2.9	0.004
P5	0.976	8.1	19.9	0.014	0.923	17.5	52.5	0.030	0.902	3.5	11.6	0.007	0.970	2.5	5.6	0.006
P6	**0.980**	6.8	18.7	0.011	0.945	12.1	33.9	0.020	**0.909**	**3.1**	**10.0**	0.004	0.919	4.2	10.2	0.004
mcMedSAM auto	0.977	7.3	19.1	0.012	0.927	12.8	38.6	0.022	0.821	4.2	15.3	0.008	0.569	3.7	9.2	0.007
MedSAM-f semi	0.999	2.4	6	0.003	0.993	3.7	12.2	0.006	0.96	2.6	8.7	0.005	0.984	2	6.9	0.003
Inter-obs	0.980	6.1	18.2	0.010	0.969	10.0	27.0	0.018	0.970	1.5	6.4	0.004	0.963	1.0	5.0	0.002

volutional architectures in LGE cardiac segmentation, they also reveal the strong dependence of the semi-automatic model on bounding-box initialization, underscoring a critical challenge for progressing toward competitive fully automatic solutions.

4.3 Why did temperature scaling outperform ensemble-based uncertainty estimation?

The temperature-scaling strategy developed by P1 clearly outperformed ensemble-based approaches for uncertainty estimation, as reflected by substantially lower CRPS scores. This superiority can be attributed to both the nature of the uncertainty in this task and the properties of the CRPS metric. Ensemble methods mainly capture epistemic uncertainty arising from model variability; however, in this challenge, most participants trained similar architectures on the same small and relatively homogeneous dataset, leading to limited ensemble diversity and consequently weak uncertainty signals. In contrast, temperature scaling directly improves probability calibration without affecting segmentation accuracy. Since CRPS rewards both calibration and sharpness, improvements in probability calibration result in more reliable confidence estimates and, consequently, lower CRPS values. Furthermore, the task is dominated by aleatoric uncertainty associated with ambiguous tissue boundaries (e.g., MI vs. MVO), for which calibration-based methods appear to be more effective than ensemble disagreement.

4.4 Can we trust automatic estimation of lesion size post-MI ?

Accurate quantification of myocardial and lesion volumes following myocardial infarction is essential for computing lesion size, expressed as the percentage of infarcted myocardium relative to the total left ventricular myocardial volume. This metric plays a critical role in diagnosis, prognosis, and therapeutic decision-making. Analysis of Tables 3 and 5 reveals that the methods developed by participants yielded relatively consistent results across both geometric and clinical metrics. Consequently, paired t-tests were performed between the MYO and MI volumes estimated by each participant and the expert reference values. To gain further insight into these results, Figure 7 displays the distributions of the estimated volumes, as well as the corresponding errors compared to reference values, for both MYO and MI volumes across participants. Four teams (P1, P3, P4, and P6) produced MYO and MI volume estimates with p-values greater than 0.05, indicating no statistically significant difference from the expert annotations. Team P2 consistently overestimated the MI volume, despite achieving top geometric scores. Conversely, P5 systematically underestimated both MYO and MI volumes. These systematic deviations likely account for the failure of the statistical tests. These findings underscore the limitations of relying solely on geometric metrics, which capture only part of the clinical relevance. Although additional validation on larger and more heterogeneous cohorts is required, the results indicate that appropriately designed deep learning methods have the potential to produce post-MI lesion size estimates

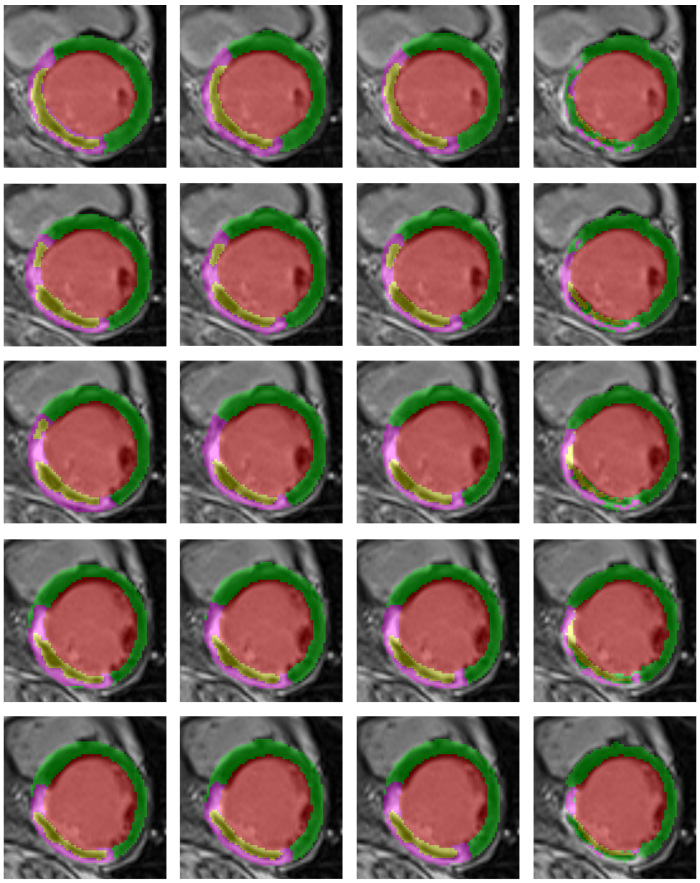


Figure 4: Representative example illustrating a typical successful case. P2 and P5 closely match the expert annotation, whereas MedSAM Auto correctly identifies the infarction and microvascular obstruction but with less accurate delineation. Left ventricular cavity (red), healthy myocardium (green), infarction (magenta), and microvascular obstruction (yellow).

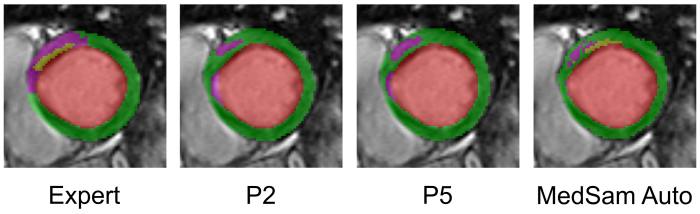


Figure 5: Representative challenging case illustrating typical failure modes. P2 and P5 confuse MVO with infarction and miss the expert-annotated infarcted region, whereas MedSAM Auto detects both structures but with localization errors. Left ventricular cavity (red), healthy myocardium (green), infarction (magenta), and microvascular obstruction (yellow).

that approach the accuracy of expert assessments.

Table 6: ASSD scores on the 3D-IR-GRE sequences from the test dataset (93% of the test data)

Method	LV	MYO	MI	MVO
P2	0.5 ±0.2	0.4 ±0.1	1.0 ±1.7	0.7 ±0.3
P5	0.6 ±0.2	0.5 ±0.2	1.2 ±1.8	1.0 ±0.5
mcMedSAM auto	0.4 ±0.1	0.4 ±0.1	1.3 ±1.2	1.9 ±1.1
Inter-obs	0.3 ±0.3	0.2 ±0.2	0.2 ±0.3	0.1 ±0.1

Table 7: ASSD scores on the 2D-PSIR sequences from the test dataset (7% of the test data). No MVO cases were present in this subset.

Method	LV	MYO	MI	MVO
P2	0.3 ±0.2	0.2 ±0.1	1.4 ±2.1	– –
P5	0.3 ±0.2	0.4 ±0.1	0.8 ±0.4	– –
mcMedSAM auto	0.3 ±0.1	0.4 ±0.1	2.2 ±1.1	– –
Inter-obs	0.2 ±0.2	0.2 ±0.1	0.1 ±0.1	– –

4.5 Generalization across sequence types

To assess the generalization capacity of the different methods, we analyzed their performance according to the acquisition type (2D PSIR vs. 3D-IR-GRE). The dataset consisted of 87% 3D and 13% 2D sequences overall, with comparable distributions in the training set (86% 3D, 14% 2D) and the test set (93% 3D, 7% 2D). Tables 6 and 7 report the ASSD scores obtained by P2, P5, and mcMedSAM-auto separately on the 3D-IR-GRE and 2D-PSIR sequences from the test dataset. These results suggest that, despite the imbalance in the training dataset with respect to the sequence type, all methods exhibit good generalization across both acquisition types. The differences in ASSD remain below 0.4 mm for all anatomical structures.

4.6 Open science initiative

In addition to the results and analyses presented in this paper on LGE cardiac MR segmentation, we support open-science principles by making the MYOSAIQ dataset available to the research community under a standard research collaboration agreement. The dataset and manual annotations

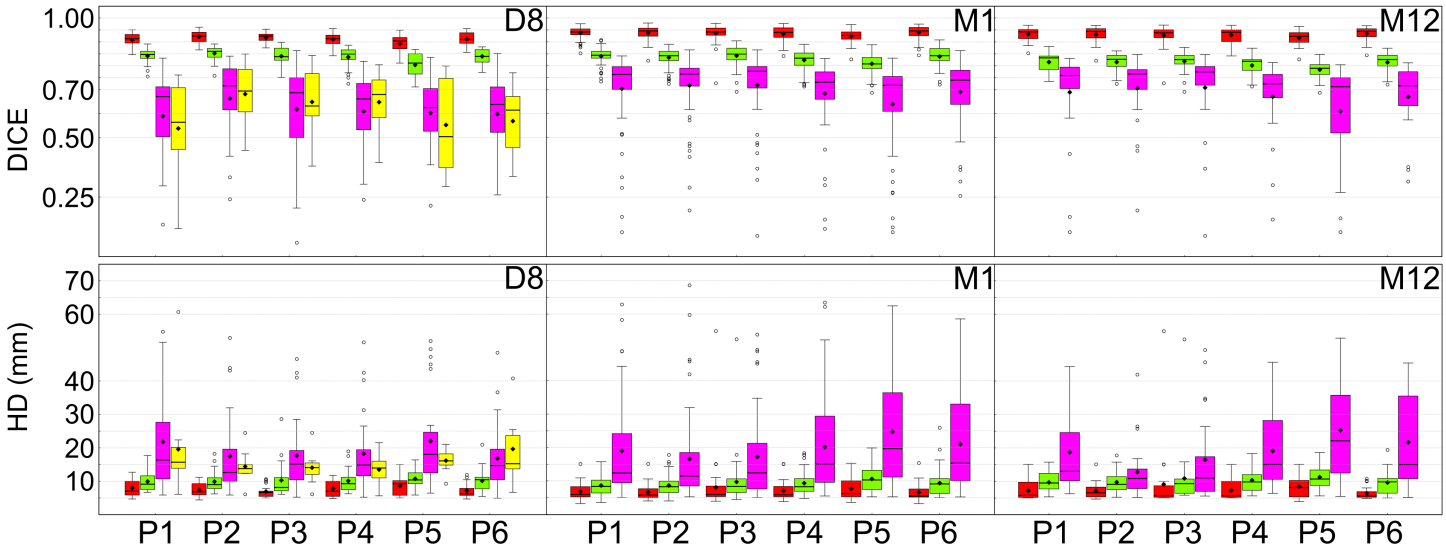


Figure 6: Distribution of Dice and Hausdorff Distance (HD) scores across participants for each label: LV cavity (red), myocardium (green), MI (magenta) and MVO (yellow). The black solid line indicates the median, while the black diamond represents the mean.

can be requested through the Human Heart Project website[2]. Although the dataset is not directly downloadable due to patient-data governance constraints, access could be granted to academic and industrial researchers upon request motivated by a research project and regulated by a specific consortium agreement. To further promote reproducibility, two participating teams (P1 and P6), which achieved among the best performances in the challenge, agreed to make their inference code publicly available. Dynamic links to the corresponding repositories are provided in Table 2. MYOSAIQ is one of the most heterogeneous datasets in LGE cardiac image analysis, encompassing major LGE imaging techniques acquired across 16 centers with three different MR vendors at three distinct post-reperfusion time points. The Codalab platform also remains open for new participants and submissions. We hope that MYOSAIQ will continue to serve as a valuable resource for the community in addressing the remaining challenges highlighted in this study.

4.7 Limitations

The definition of a gold standard based on expert annotations remains uncertain. Researchers in this applicative field still debate on which reference methods offer the most reliable baseline in clinical practice, whether 5σ or 6σ thresholds compared to remote-myocardium mean values, or approaches such as Full Width at Half Maximum (Karim et al., 2016b). There is also no established methodology for quality assurance or final validation when segmentations come from several experts working on LGE images. Because the absolute accuracy of any segmentation, including those made by experts, cannot be measured, we can only report reproducibility and inter-observer variation limited to two experts in this study. In our work, disagreements between observers were not resolved through consensus or arbitration, particularly at lesion borders where opinions tend to diverge. The inter-observer variability shown here should therefore be viewed as a worst-case estimate. The results should also be interpreted with caution because they are based on submissions from only six participating teams. The number of samples in the dataset is below 500, which restricts the strength of the conclusions and may not reflect the broader state of the art. Finally, the MYOSAIQ multi-centre database combines two studies conducted on scanners available at the time. Although it captures a wide range of realistic clinical patterns, it mirrors the scanner market distribution of that period and does not include newer or less common systems. It cannot guarantee fairness across vendors or scanner types, nor does it ensure gender balance. These factors limit how far algorithms trained for this challenge can be generalized. Still, the MYOSAIQ datasets will remain available for future projects that may pool more data and support the development of shared, larger-scale solutions.

4.8 Conclusions

This paper presents a comprehensive study on the automatic quantification of myocardial infarction using a multi-center, multi-vendor LGE cardiac MRI dataset. Results were obtained from six teams participating in the MYOSAIQ challenge, as well as from foundation models specifically designed to address the generalization challenge in medi-

2. https://humanheart-project.creatis.insa-lyon.fr/database/#collections

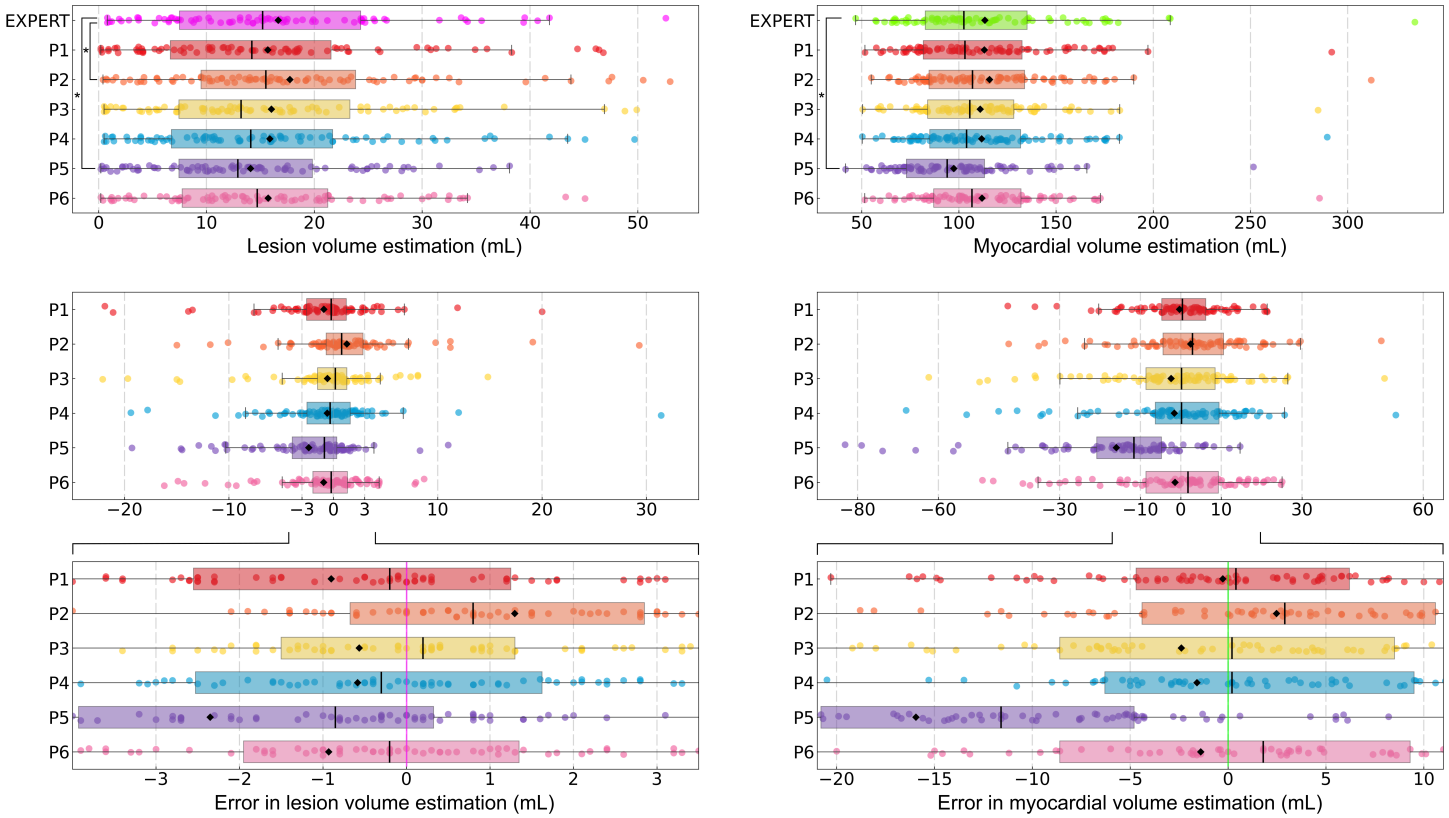


Figure 7: Estimation of lesion (MI) and myocardial (MYO) volumes, along with the distribution of estimation errors across participants. The black line indicates the median, while the black diamond denotes the mean. Paired t-tests were conducted to compare each participant's volume estimation with the reference derived from expert segmentation. Brackets marked with an asterisk indicate statistically significant differences ($p < 0.05$).

cal imaging. Inter-observer variability was also assessed to benchmark the performance of each method. This study confirms that LV segmentation achieves high levels of performance under controlled conditions, while varying levels of difficulty persist for MYO, MI, and MVO, with MI and MVO being the most challenging. The comparison across participants suggests that UNet-inspired backbone architectures may reach a performance plateau, with limited benefit from the evaluated data augmentation strategy when relying solely on image data. Additionally, fine-tuned foundation models appear to have the potential to overcome these limitations, likely due to encoders that have learned generic features, enabling simple yet effective fine-tuning of lightweight decoders. In terms of clinical metrics, deep learning methods achieved competitive correlation, MAE, and LOA scores for LV and MYO volumes comparable to expert variability. Regarding MI lesion size, four out of six teams produced volume estimates that were not significantly different from the expert's reference values, suggesting the feasibility of automatic measurement of this index in clinical routine.

Acknowledgments

This work was supported in part by: the RHU MARVELOUS (ANR-16-RHUS-0009) of l'Université Claude Bernard Lyon 1 (UCBL), within the program "Investissements d'Avenir" operated by the French National Research Agency (ANR). We also thank Olea Medical and the Virtual Physiological Human Institute (VPHi) for sponsoring the awards for the MYOSAIQ challenge at the FIMH 2023 conference. We also thank the support by AEI (Spain) through projects TED2021-130459B-I00 and PID2022-140556OB-I00 and Aragón Gov. through T39_23R.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

The MYOSAIQ dataset is available to the research community under a standard research collaboration agreement. The dataset and manual annotations can be requested through the Human Heart Project website[3]. Although the dataset is not directly downloadable due to patient-data governance constraints, access could be granted to academic and industrial researchers upon request motivated by a research project and regulated by a specific consortium agreement.

References

Luciano C Amado, Bernhard L Gerber, Sandeep N Gupta, Dan W Rettmann, Gilberto Szarf, Robert Schock, Khurram Nasir, Dara L Kraitchman, and João AC Lima. Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model. *Journal of the American College of Cardiology*, 44(12):2383–2389, 2004.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.04517.

Loic Belle, Pascal Motreff, Lionel Mangin, Grégoire Rangé, Xavier Marcaggi, Antoine Marie, Nadine Ferrier, Olivier Dubreuil, Gilles Zemour, Géraud Souteyrand, et al. Comparison of immediate with delayed stenting using the minimalist immediate mechanical intervention approach in acute st-segment–elevation myocardial infarction: The mimi study. *Circulation: Cardiovascular Interventions*, 9(3):e003388, 2016.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. .

Heerajnarain Bulluck, Stefania Rosmini, Amna Abdel-Gadir, Anish N Bhuva, Thomas A Treibel, Marianna Fontana, Shane Weinmann, Alex Sirker, Anna S Herrey, Charlotte Manisty, James C Moon, and Derek J Hausenloy. Impact of microvascular obstruction on semiautomated techniques for quantifying acute and chronic myocardial infarction by cardiovascular magnetic resonance. *Open Heart*, 3: e000535, 2016.

Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7 (25), 2020. .

3. http://humanheart-project.creatis.insa-lyon.fr/

Nicolas Duchateau, Magalie Viallon, Lorena Petrusca, Patrick Clarysse, Nathan Mewton, Loic Belle, and Pierre Croisille. Pixel-wise statistical analysis of myocardial injury in stemi patients with delayed enhancement mri. *Frontiers in Cardiovascular Medicine*, 10:1136760, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Celia Goujat, Pierre-Marc Jodoin, Loïc Boussel, and Olivier Bernard. Medical sam for lge-mri cardiac segmentation: promise or hype? In *Statistical Atlases and Computational Modeling of the Heart–STACOM 2025: Daejeon, Republic of Korea, September 27, 2025, proceedings*. Springer, 2025.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Borja Ibanez, Anthony H. Aletras, Andrew E. Arai, Hakan Arheden, Jeroen Bax, Colin Berry, Chiara Bucciarelli-Ducci, Pierre Croisille, Erica Dall'Armellina, Rohan Dharmakumar, Ingo Eitel, Rodrigo Fernández-Jiménez, Matthias G. Friedrich, David García-Dorado, Derek J. Hausenloy, Raymond J. Kim, Sebastian Kozerke, Christopher M. Kramer, Michael Salerno, Javier Sánchez-González, Javier Sanz, and Valentin Fuster. Cardiac mri endpoints in myocardial infarction experimental and clinical trials: Jacc scientific expert panel. *Journal of the American College of Cardiology*, 74(2):238–256, 2019. ISSN 0735-1097. . URL https://www.sciencedirect.com/science/article/pii/S0735109719352921.

Rashed Karim, Pranav Bhagirath, Piet Claus, R James Housden, Zhong Chen, Zahra Karimaghaloo, Hyon-Mok Sohn, Laura Lara Rodríguez, Sergio Vera, Xènia Albà, Anja Hennemuth, Heinz-Otto Peitgen, Tal Arbel, Miguel A Gonzàlez Ballester, Alejandro F Frangi, Marco Götte, Reza Razavi, Tobias Schaeffter, and Kawal Rhode. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement MR images. *Med Image Anal*, 30:95–107, January 2016a.

Rashed Karim, Pranav Bhagirath, Piet Claus, R. James Housden, Zhong Chen, Zahra Karimaghaloo, Hyon-Mok Sohn, Laura Lara Rodríguez, Sergio Vera, Xènia Albà, Anja Hennemuth, Heinz-Otto Peitgen, Tal Arbel, Miguel A. Gonzàlez Ballester, Alejandro F. Frangi, Marco Götte, Reza Razavi, Tobias Schaeffter, and Kawal Rhode. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement mr images. *Medical Image Analysis*, 30:95–107, 2016b. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841516000050.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

Alain Lalande, Zhihao Chen, Thibaut Pommier, Thomas Decourselle, Abdul Qayyum, Michel Salomon, Dominique Ginhac, Youssef Skandarani, Arnaud Boucher, Khawla Brahim, Marleen de Bruijne, Robin Camarasa, Teresa M. Correia, Xue Feng, Kibrom B. Girum, Anja Hennemuth, Markus Huellebrand, Raabid Hussain, Matthias Ivantsits, Jun Ma, Craig Meyer, Rishabh Sharma, Jixi Shi, Nikolaos V. Tsekos, Marta Varela, Xiyue Wang, Sen Yang, Hannu Zhang, Yichi Zhang, Yuncheng Zhou, Xiahai Zhuang, Raphael Couturier, and Fabrice Meriaudeau. Deep learning methods for

automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. *Medical Image Analysis*, 79:102428, 2022. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841522000792.

Erwan Lecesne, Antoine Simon, Mireille Garreau, Gilles Barone-Rochette, and Céline Fouard. Segmentation of cardiac infarction in delayed-enhancement mri using probability map and transformers-based neural networks. *Computer Methods and Programs in Biomedicine*, 242:107841, 2023. ISSN 0169-2607. . URL https://www.sciencedirect.com/science/article/pii/S0169260723005072.

Lei Li, Fuping Wu, Sihan Wang, Xinzhe Luo, Carlos Martín-Isla, Shuwei Zhai, Jianpeng Zhang, Yanfei Liu, Zhen Zhang, Markus J. Ankenbrand, Haochuan Jiang, Xiaoran Zhang, Linhong Wang, Tewodros Weldebirhan Arega, Elif Altunok, Zhou Zhao, Feiyan Li, Jun Ma, Xiaoping Yang, Elodie Puybareau, Ilkay Oksuz, Stephanie Bricq, Weisheng Li, Kumaradevan Punithakumar, Sotirios A. Tsaftaris, Laura M. Schreiber, Mingjing Yang, Guocai Liu, Yong Xia, Guotai Wang, Sergio Escalera, and Xiahai Zhuang. Myops: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. *Medical Image Analysis*, 87:102808, 2023. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841523000695.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15 (1):654, Jan 2024. ISSN 2041-1723. . URL https://doi.org/10.1038/s41467-024-44824-z.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023. URL http://jmlr.org/papers/v24/21-1436.html.

S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P.F. Jaeger, and K. Maier-Hein. Mednext: Transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023.

Magalie Viallon, Alexis Jacquier, Carmen Rotaru, Bénédicte MA Delattre, Nathan Mewton, Fabrice Vincent, and Pierre Croisille. Head-to-head comparison of eight late gadolinium-enhanced cardiac mr (lge cmr) sequences at 1.5 tesla: from bench to bedside. *Journal of magnetic resonance imaging*, 34(6):1374–1387, 2011a.

Magalie Viallon, Alexis Jacquier, Carmen Rotaru, Bénédicte M.A. Delattre, Nathan Mewton, Fabrice Vincent, and Pierre Croisille. Head-to-head comparison of eight late gadolinium-enhanced cardiac mr (lge cmr) sequences at 1.5 tesla: From bench to bedside. *Journal of Magnetic Resonance Imaging*, 34(6):1374–1387, 2011b. . URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.22783.

Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*, 23(7):903–921, July 2004.

Derek M. Yellon, David J. Hearse, Renata Crome, and Richard K.H. Wyse. Temporal and spatial characteristics of evolving cell injury during regional myocardial ischemia in the dog: The "border zone" controversy. *Journal of the American College of Cardiology*, 2(4):661–670, 1983. ISSN 0735-1097. . URL https://www.sciencedirect.com/science/article/pii/S0735109783803064.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2020. .