# Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

Use the title to convey the essential information on the challenge mission.

Automatic Evaluation of Myocardial Infarction from Delayed-Enhancement Cardiac MRI

### Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

EMIDEC

### Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

One crucial parameter to evaluate the state of the heart after myocardial infarction (MI) is the viability of the myocardial segment, i.e. if the segment recovers its functionality upon revascularization. MRI performed several minutes after the injection of a contrast agent (delayed enhancement-MRI or DE-MRI) is a method of choice to evaluate the extent of MI, and by extension, to assess viable tissues after an injury (in conjunction with the thickening of the muscle evaluated from cine-MRI) [3]. The two main objectives of this challenge are first to classify normal and pathological cases from the clinical information with or without DE-MRI, and secondly to automatically detect the different relevant areas (the myocardial contours, the infarcted area and the permanent microvascular obstruction area (no-reflow area)) from a series of short-axis DE-MRI covering the left ventricle. The segmentation allows us to make a quantification of the MI, in absolute value (mm3) or percentage of the myocardium.
The database consists of 150 exams (all from different patients) divided into 50 cases with normal MRI after injection of a contrast agent and 100 cases with myocardial infarction (and then with a hyperenhanced area on DE-MRI), whatever their inclusion in the cardiac emergency department. Along with MRI, clinical characteristics are provided to distinguish normal and pathological cases. The training set (100 cases) as a dedicated online website will be available mid-April. To participate to the challenge and get access to the datasets, the participant should create an account through this dedicated online evaluation website. Moreover, the participants will also be requested to submit a paper following the MICCAI format that describe the methodology. The submitted papers will be accepted after a deep proofreading.
The segmentation contest will take place in July, then before the conference, and the global ranking will be based on geometrical and clinical metrics currently used in medical practices. The classification contest will happen during the conference, and the global ranking will correspond to the classification accuracy.

### Challenge keywords

List the primary keywords that characterize the challenge.

MRI, heart, myocardial infarction, segmentation, classification, delayed-enhancement

### Year

The challenge will take place in ...

2020

## FURTHER INFORMATION FOR MICCAI ORGANIZERS

### Workshop

If the challenge is part of a workshop, please indicate the workshop.

Stacom

### Duration

How long does the challenge take?

Half day.

### Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect a relatively large number of participants, somewhere in between 15 to 20 research teams. As a comparison, the ACDC challenge in the MICCAI conference in 2017 received results from 10 different methods [1]. A preliminary challenge organized in 2012 by Karim et al. was already dedicated to the automatic processing of DE-MRI [2]. Five groups took part in this challenge. The results were very interesting, but there are some differences with our proposal. First, the dataset in 2012 was composed of fifteen human and fifteen porcine datasets, all pathological cases. There was no associated clinical data.
Now, there are more and more laboratories working on the automatic post-processing of cardiac MRI and we expect a strong participation for this challenge. One of the main advantages of our database is the associated clinical data, simulating the classic workflow in emergency services.

### Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

One publication in Q1 journal summarizing all the results of the challenge (Q1 refer to the first quartile of the journal ranking quartiles within a sub-discipline using the SJR citation index (https://www.scimagojr.com/)). The website will stay open after the challenge at least 5 years to enable new submissions and evaluations as well as an update in the leaderboard. The papers will be downloadable for the next 5 years. The data will be available after the conclusion of the challenge. The data will be uploaded to a repository such as Zenodo and therefore downloadable as well. To be more precise, the data available after the conclusion of the challenge is the entire training dataset (images, clinical information and ground-truths), as well as the images and the associated clinical information for the testing dataset. The testing dataset is different between the segmentation contest and the classification contest.

## Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

Partially online challenge. A projector and a microphone would be appreciated if the room is large. A conventional computer is required to do the ranking in real time of the classification contest.

# TASK: Segmentation contest

## SUMMARY

### Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Segmentation of the myocardium for all the DE-MRI exams, and the myocardial infarction and no-reflow areas on exams classified as pathologic ones. Contest done before the conference.

### Keywords

List the primary keywords that characterize the task.

MRI, heart, myocardial infarction, normal case, delayed-enhancement, no-reflow, segmentation

## ORGANIZATION

### Organizers

a) Provide information on the organizing team (names and affiliations).

Alain Lalande, ImVia Laboratory and University Hospital of Dijon, Dijon, France (alain.lalande@u-bourgogne.fr)
Fabrice Meriaudeau, ImVia Laboratory, Dijon, France (fabrice.meriaudeau@u-bourgogne.fr)
Dominique Ginhac, ImVia Laboratory, Dijon, France (dominique.ginhac@ubfc.fr)
Abdul Qayyum, ImVia Laboratory, Dijon, France (abdul.qayyum@u-bourgogne.fr)
Khawla Brahim, ImVia Laboratory, Dijon, France (khawla.enim@gmail.com)
Thibaut Pommier, University Hospital of Dijon, Dijon, France (thibaut.pommier@chu-dijon.fr)
Raphaël Couturier, Femto-ST laboratory, Belfort, France (raphael.couturier@univ-fcomte.fr)
Michel Salomon, Femto-ST laboratory, Belfort, France (michel.salomon@univ-fcomte.fr)
Gilles Perrot, Femto-ST laboratory, Belfort, France (gilles.perrot@univ-fcomte.fr)
Zhihao Chen, Femto-ST laboratory, Belfort, France (zhihao.chen@femto-st.fr)

b) Provide information on the primary contact person.

Alain Lalande, ImVia Laboratory and University Hospital of Dijon, Dijon, France (alain.lalande@u-bourgogne.fr)

### Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

One time event.

## Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

A specific website will be hosted at grand-challenge.org. The website will describe precisely the challenge including a general presentation of the challenge, of the organizing committee, the important dates, task presentation, data description and rule presentation. The results will be published on this website also after the conference.

c) Provide the URL for the challenge website (if any).

Work in progress, it would be ready at the beginning of February.

## Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic., Semi automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Data will be made 100% public without the need for the participants to fill out any copyright form. There are no restrictions with respect to usage of other data.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Challenge prize will be symbolic.
Award for the team classified as first.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Ranking of all the teams according to the evaluation parameters.
The different metrics are automatically calculated. The retained metrics will be provided for each team after the challenge (in August 2020), and the global ranking will be provided during the conference.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

We plan to do a survey paper that shall describe the dataset as well as the ground-truthing and validation

processes, and report in great details results obtained by each participant. The participating teams will be invited to contribute to this journal paper. A maximum of two members of the participating teams will be qualifyed as co-authors. The other co-authors will be the organizers of the challenge. The paper will be submitted to a high impact journal in the field. The organizers will review the paper for sufficient details to be able to understand and reproduce the method and hold the right to exclude participants from the joint journal paper in case their method description is not adequate.
This article will compile the results of task 1 and task 2.

## Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

This challenge will take place before the conference. The participants are invited to submit their results (corresponding to the segmentation of the different areas) through the dedicated website by loading a file in Nifti format with all the areas of interest. The algorithm will not be requested. Each participant is allowed to submit up to three times his results, the organizers keeping only the best performing model among the three attempts. Moreover, the participants will be requested to submit an article of four pages, following the MICCAI format, describing the methodology. The submitted papers will be accepted after a deep proofreading. Articles will be reviewed by the organizing board and published on line if meeting the expected quality.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

During the training phase, the code for the metric calculation will be available to the participants and thus they can assess themselves their method. Specific information will be provided via the dedicated website in order to describe the two different tasks (segmentation contest and classification contest), including dataset information and access. After a registration step, the participants can download the training dataset, i.e. images (in Nifti format) and the ground-truths (also in Nifti format).

During the testing phase, each participant can submit only three times their results. The best one will be retained. The retained metrics will be provided for each team after the challenge (in August 2020). The global ranking will be provided after the challenge, during the conference.

## Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Releases
Mid-April : Release of the training cases

Mid-July and for 2 weeks: Release of the testing dataset

Registration and challenge
Mid-April : Start of the registration process
Mid of July: End of registration
End of July: End of the challenge (deadline for the submission of the paper)
Mid-August: Results of the challenge

## Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

The overall dataset was created from real clinical exams acquired at the University Hospital of Dijon (France). Acquired data were fully anonymized and handled within the regulations set by the local ethical committee. As the data were collected retrospectively, and as the data are completly untraceable (because using the NifTI format, we discard all the administrative information included in the header), for the french law, and for the staff of the ethical committee of the University Hospital of Dijon, it was not necessary to do the process to have ethical approval number. The ethical committee of University Hospital of Dijon will check the compliance with the law of the created dataset.

## Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC SA.

Additional comments: The associated publication summarizing the challenge and the University Hospital must be cited in any publication. Moreover, participants may use other datasets for the development of a method that will be submitted to the challenge, provided that the datasets are publicly available and clearly stated in the submitted paper.

## Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Creation of the website hosted at grand-challenge.org in progress. Then the code for the evaluation and the metric calculation will be directly available via this website.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

No code request from the participating teams.

### Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No conflict of interest.

## MISSION OF THE CHALLENGE

### Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Diagnosis, Decision support, Prognosis.

### Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation.

## Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The targeted cohort is any patient admitted in cardiac emergency department with symptoms of heart attack. A MRI of the left ventricle (in short axis orientation) acquired several minutes after the injection of a contrast agent (delayed enhancement-MRI or DE-MRI) allows for evaluating the extent of myocardial infarction.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The cohort consists of data extracted from 150 MRI exams (all from different patients) divided into 50 cases with normal MRI after the injection of a contrast agent and 100 cases with myocardial infarction (and then with a hyperenhanced area on DE-MRI), whatever their inclusion in the cardiac emergency department. The cases were randomly selected from our database. The inclusion criteria are patients received in the cardiac emergency department with acute disease (with symptoms of heart attack) and that undergo cardiac MRI. The exclusion criteria are patients with contraindications to the MRI and cardiac chronic diseases. There is an unbalanced distribution between normal and pathological cases, corresponding roughly to real life in managed exams in a MRI department. The overall dataset was created from real clinical exams acquired from the MRI department at the University Hospital of Dijon (France). Each group was clearly defined according to physiological parameters and the presence or absence of a disease area on DE-MRI. The data are DE-MRI in short axis orientation, and a series of images covering the left ventricle.

## Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

MRI (1.5 T and 3 T scanners)

## Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Spatial resolution.

b) ... to the patient in general (e.g. sex, medical history).

No clinical information, because these data are used for task 2.

## Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

MRI of the left ventricle of the heart in short axis orientation after the injection of a gadolinium-based contrast agent.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm target is the myocardium, as the infarcted area and the no-reflow area, if present.

### Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Robustness, Reliability, Precision, Accuracy.

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Siemens MRI scanners (Area (1.5 T) and Skyra (3T))

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Conventional cardiovascular exam. No specific protocol. Retrospective study where we extract only the short-axis slices of the DE-MRI.
In details, all acquisitions are ECG-gated, taken during breath-hold and performed 10 minutes after the injection of a gadolinium-based contrast agent. A T1-weighted Phase Sensitive Inversion Recovery (PSIR) sequence is used. Resulting MR images consist in a stack of short-axis slices from base to apex of the left ventricle with the following features: pixel spacing between 1.25 × 1.25 $mm^2$ and 2 × 2 $mm^2$, slice thickness of 8 mm and distance between slices of 10 mm (i.e. one image every 10 mm). The variation of these parameters at the acquisition on 1.5 T or 3 T magnets allows us to deal with images with different signal to noise ratio. To prevent the drawback of the displacement of the heart location between slices due to different breath-holds, the slices are realigned according to the gravity center of the area defined by the epicardial contour. The raw input images will be provided using Nifti format, i.e. one file for the whole images covering the left ventricle, and one file with the ground-truths.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

University Hospital of Dijon (France)

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

A cardiologist with 10 years of experience in cardiology and MRI and a biophysicist with 20 years of experience of cardiovascular MRI supervise the data acquisition.

### Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Every training and test case represents a DE-MRI exam of the left ventricle. An exam (i.e. a case) consists of a series of short-axis slices covering the left ventricle from the base to the apex. The ground-truths (contours of the relevant areas) will be provided with the training dataset.

b) State the total number of training, validation and test cases.

100 cases for the training and 50 cases for the testing. From 5 to 10 slices per exam, covering the left ventricle. The training set with full ground-truth will comprise 100 cases (67 pathological cases, 33 normal cases) randomly selected among the 150 subjects. The testing set is made of data from 50 subjects (33 pathological cases, 17 normal cases), all different from those in the training set.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Classic proportion of training and test cases for a challenge such as the ACDC challenge [1].

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Unbalanced distribution between normal (1/3) and pathological (2/3) cases, corresponding roughly to real life in managed exams in a  MRI department.

### Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

The contours are manually drawn for all the cases. The gold standard for the contours is obtained through a manual segmentation carried out by two experts. The left ventricular endocardial and epicardial borders, as well as the infarcted area and the no reflow areas, if present, were first outlined by the first expert, an experienced user (a cardiologist with 10 years of experience in cardiology and MRI). Then, the second expert (a well-trained

biophysicist with 20 years of experience) went through every outline and made some changes when necessary. These contours are then transcribed in label field image. From the contours, specific labels are assigned to each voxel depending on their location: in the background, in the myocardium, in the myocardial infarction area and in the no-reflow area, respectively.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

A specific rule for the myocardial segmentation: the papillary muscles are included in the cavity.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

The different areas are first outlined by an experienced user (a cardiologist with 10 years of experience in cardiology and MRI) and then the segmentation is verified by a second expert (a well-trained biophysicist with 20 years of experience). Same process for the training and test cases.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

The first expert outlines the contours, and the second goes through every outline and carries out some changes if necessary.

### Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

To prevent the drawback of the displacement of the heart location between slices due to different breath-holds, the slices are realigned according to the gravity center of the area defined by the epicardial contour. Same process for the training and test cases.

### Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Low contrast between the normal myocardium and the surrounding areas, such as the lung or the liver.
Low contrast between the cavity and the myocardial infarction.
No-reflow areas are very small.
Some artefacts due to the acquisition could hamper the segmentation of the myocardial infarction.
In our own experience, the intra-observer variability is around 3% for the myocardium, and 5% for the disease areas, as 5% for the myocardium and 7% for the disease areas for the inter-observer variability.

b) In an analogous manner, describe and quantify other relevant sources of error.

No other sources of error. The potential displacement of the heart in the image from one slice to another, due to different breath-holds is corrected from the gravity center of the area defined by the epicardial contour.

## ASSESSMENT METHODS

## Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

The clinical metrics are those that are the most widely used in cardiac clinical practice i.e. the average errors for the volume of the left ventricle (in mm3), the volume (in mm3) and the percentage of MI and no-reflow. The geometrical metrics are the average DICE index and Hausdorff distance (in 3D) for the different areas [4]. In our opinion, it is not necessary to extend the evaluation parameters, and we focus to keep the most relevant ones, without any bias or link between them.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The clinical metrics simulate the needed information for the medical doctor, and the geometrical metrics are the classic ones used in the segmentation evaluation. The Dice index gives an overall information about the quality of the segmentation, the Hausdorff distance highlights the outliers [4]. In our opinion, it is not necessary to extend the evaluation parameters, and we focus on keeping the more relevant ones, without any bias or link between them.

## Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Ranking will be based on geometrical  and clinical metrics currently used in medical practices. Then for each metric, a ranking will be done, and the final ranking consists of the sum of the ranking for each metric. By limiting the bias and link between the evaluation parameters, doing the sum of them makes more sense.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Participant teams must provide results for all the cases, otherwise, they will not appear in the final ranking.

c) Justify why the described ranking scheme(s) was/were used.

Doing the sum of the ranking of few metrics (with no bias) is simple and easy to understand. The limited number of metrics is a deliberate choice.

## Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

No missing data handling.
No specific statistical methods, except the ranking

b) Justify why the described statistical method(s) was/were used.

No need to have a deep statistical evaluation.

## Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

No further analysis.

# TASK: Classification contest

## SUMMARY

### Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Classify the exams in normal or pathological one, according to the clinical data with or without the DE-MRI exams (two sub-challenges, the first one from only the clinical informations, and the second one considering the clinical informations and the DE-MRI). Contest done online and on-site.
Some cases could be sometimes ambiguous rendering this task not evident. Indeed patients coming in an emergency department could have other diseases, providing normal DE-MRI but abnormals clinical informations. For example, myocarditis could provide abnormal values for some clinical parameters, but normal DE-MRI. However, even if one parameter is ambiguous, considering the whole provided clinical parameters will prevent any big ambiguity.

### Keywords

List the primary keywords that characterize the task.

MRI, heart, myocardial infarction, normal case, delayed-enhancement, classification

## ORGANIZATION

### Organizers

a) Provide information on the organizing team (names and affiliations).

Alain Lalande, ImVia Laboratory and University Hospital of Dijon, Dijon, France (alain.lalande@u-bourgogne.fr)
Fabrice Meriaudeau, ImVia Laboratory, Dijon, France (fabrice.meriaudeau@u-bourgogne.fr)
Alexandre Cochet, ImVia Laboratory, Dijon, France (alexandre.cochet@u-bourgogne.fr)
Dominique Ginhac,  ImVia Laboratory, Dijon, France (dominique.ginhac@ubfc.fr)
Thibaut Pommier, CHU Dijon, Dijon, France (thibaut.pommier@chu-dijon.fr)
Raphaël Couturier, Femto-ST laboratory, Belfort, France (raphael.couturier@univ-fcomte.fr)

b) Provide information on the primary contact person.

Alain Lalande, ImVia Laboratory and University Hospital of Dijon, Dijon, France (alain.lalande@u-bourgogne.fr)

### Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

One time event.

## Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

A specific website will be hosted at grand-challenge.org. The website will describe precisely the challenge including a general presentation of the challenge, of the organizing committee, the important dates, task presentation, data description and rule presentation. The results will be published on this website also after the conference.

c) Provide the URL for the challenge website (if any).

Work in progress,  it would be ready at the beginning of February.

## Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Data will be made 100% public without the need for the participants to fill out any copyright form. There are no restrictions with respect to usage of other data.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Challenge prize will be symbolic.
Award for the teams classified as first for each sub-challenge (two sub-challenges).

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Ranking of all the teams according to the classification accuracy.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

We plan to do a survey paper that shall describe the dataset as well as the ground-truthing and validation processes, and report in great details results obtained by each participant. The participating teams will be invited

to contribute to this journal paper. A maximum of two members of the participating teams will be qualifyed as co-authors. The other co-authors will be the organizers of the challenge. The paper will be submitted to a high impact journal in the field. The organizers will review the paper for sufficient details to be able to understand and reproduce the method and hold the right to exclude participants from the joint journal paper in case their method description is not adequate.
This article will compile the results of task 1 and task 2.

## Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

On-site challenge. The participants will have one hour to run their methods on their own laptop in order to classify the exams between normal and pathologic cases. They will provide a text file with the results of their classification (binary classification). Then a ranking of the methods based on the classification accuracy will be provided immediately during the challenge. The algorithm will not be requested. Moreover, the participants will be requested to submit an article of four pages before the conference, following the MICCAI format,describing the methodology. The submitted papers will be accepted after a deep proofreading. Articles will be reviewed by the organizing board and published on line if meeting the expected quality.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

During the training phase, the participants can directly evaluate the accuracy of their approach thanks to the ground-truth. Specific information will be provided via the dedicated website in order to describe the two different tasks (segmentation contest and classification contest), including dataset information and access. After a registration step, the participants can download the training dataset, i.e. images (in Nifti format) and ground-truths (also in Nifti format), and the clinical information (one text file for each exam).

## Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Releases
Mid-April : Release of the training cases
Online and on-site challenge (release of the testing cases during the conference)

Registration and challenge
Mid-April: Start of the registration process
End of August:  End of registration (deadline for the submission of the paper)

Online and on-site challenge

## Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

The overall dataset was created from real clinical exams acquired at the University Hospital of Dijon (France). Acquired data were fully anonymized and handled within the regulations set by the local ethical committee. As the data were collected retrospectively, and as the data are completly untraceable (because using the NifTI format, we discard all the administrative information included in the header), for the French law, and for the staff of the ethical committee of the University Hospital of Dijon, it was not necessary to undergo the process of applying for an ethical approval number.
The clinical information are not specific enough to retrieve a specific patient.
The ethical committee of University Hospital of Dijon will check the compliance with the law of the created dataset.

## Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC SA.

Additional comments: The associated publication summarizing the challenge and the University Hospital must be cited in any publication. Moreover, participants may use other datasets for the development of a method that will be submitted to the challenge, provided that the datasets are publicly available and clearly stated in the submitted paper.

## Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Creation of the website hosted at grand-challenge.org in progress. The ground-truth available via this website will allow the participants to evaluate the accuracy of their method themselves.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

No code request from the participating teams.

### Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No conflict of interest.

## MISSION OF THE CHALLENGE

### Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Diagnosis, Decision support.

### Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification.

### Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The targeted cohort is any patient admitted in cardiac emergency department with symptoms of heart attack. A MRI of the left ventricle (in short axis orientation) acquired several minutes after the injection of a contrast agent (delayed enhancement-MRI or DE-MRI) allows to differentiate the case in one sub-challenge.
Additionnal clinical information available during the management of the patient in a clinical emergency department should reinforce this classification.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The cohort consists of data extracted from 150 MRI exams (all from different patients) divided into 50 cases with normal MRI after injection of a contrast agent and 100 cases with myocardial infarction (and then with a hyperenhanced area on DE-MRI), whatever their inclusion in the cardiac emergency department. The cases were randomly selected from our database. The inclusion criteria are patients received in the cardiac emergency department with acute disease (with symptoms of heart attack) and that undergo cardiac MRI. The exclusion criteria are patients with contraindications to the MRI and cardiac chronic diseases. There is an unbalanced distribution between normal and pathological cases, corresponding roughly to real life in managed exams in a MRI department. The overall dataset was created from real clinical exams acquired from the MRI department at the University Hospital of Dijon (France). Each group was clearly defined according to physiological parameters and the presence or absence of a disease area on DE-MRI. The data are DE-MRI in short axis orientation, and a series of images covering the left ventricle.
Along with MRI, clinical characteristics are provided to distinguish normal and pathological cases. These characteristics are: sex, age, tobacco (Y/N), overweight (BMI > 25), arterial hypertension (Y/N), diabetes (Y/N), familial history of coronary artery disease (Y/N), ECG (ST+ (STEMI) or not), troponin (value), Killip max (between 1 and 4), ejection fraction of the left ventricle from echography (value), NTproBNP (value) and SYNTAX score (value).

### Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

MRI (1.5 T and 3 T scanners)

### Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Spatial resolution.

b) ... to the patient in general (e.g. sex, medical history).

The following clinical information are provided : sex, age, tobacco (Y/N), overweight (BMI > 25), arterial hypertension (Y/N), diabetes (Y/N), familial history of coronary artery disease (Y/N), ECG (ST+ (STEMI) or not), troponin (value), Killip max (between 1 and 4), ejection fraction of the left ventricle from echography (value), NTproBNP (value) and SYNTAX score (value).

### Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

MRI of the left ventricle of the heart in short axis orientation after the injection of a gadolinium-based contrast agent.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

The algorithm target is the classification of each exam as normal or pathological.

### Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Siemens MRI scanners (Area (1.5 T) and Skyra (3T))

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

- Conventional cardiovascular exam. No specific protocol. Retrospective study where we extract only the short-axis slices of the DE-MRI.
All acquisitions are ECG-gated, taken during breath-hold and performed 10 minutes after the injection of a gadolinium-based contrast agent. A T1-weighted Phase Sensitive Inversion Recovery (PSIR) sequence is used. Resulting MR images consist in a stack of short-axis slices from base to apex of the left ventricle with the following features: pixel spacing between 1.25 × 1.25 mm² and 2 × 2 mm², slice thickness of 8 mm and distance between slices of 10 mm (i.e. one image every 10 mm). The variation of these parameters at the acquisition on 1.5 T or 3 T magnets allows us to deal with images with different signal to noise ratio. To prevent the drawback of the displacement of the heart location between slices due to different breath-holds, the slices are realigned according to the gravity center of the area defined by the epicardial contour. The raw input images will be provided using

Nifti format, i.e. one file for the whole images covering the left ventricle, and one file with the ground-truths.
- Classic clinical information recorded during the management of the patient. No specific protocol.
These clinical information are : sex, age, tobacco (Y/N), overweight (BMI > 25), arterial hypertension (Y/N), diabetes (Y/N), familial history of coronary artery disease (Y/N), ECG (ST+ (STEMI) or not), troponin (value), Killip max (between 1 and 4), ejection fraction of the left ventricle from echography (value), NTproBNP (value) and SYNTAX score (value). These clinical information will be provided in a text file.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

University Hospital of Dijon (France)

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

A cardiologist with 10 years of experience in cardiology and MRI and a biophysicist with 20 years of experience of cardiovascular MRI supervise the data management.

### Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Every training and test case represents a DE-MRI exam of the left ventricle. An exam (i.e. a case) consists of a series of short-axis slices covering the left ventricle from the base to the apex. The ground-truths (contours of the relevant areas) will be provided with the training dataset.
For each exam, associated clinical information will be provided in a text file.

b) State the total number of training, validation and test cases.

100 cases for the training and 50 cases for the testing. From 5 to 10 slices per exam, covering the left ventricle.
The training set with full ground-truth will comprise 100 cases (67 pathological cases, 33 normal cases) randomly selected among the 150 subjects.
The testing set is made of data from 50 subjects (33 pathological cases, 17 normal cases), all different from those in the training set. The images and the clinical information will be provided at the beginning of the challenge. Two sub-contests are designed, the first one consists in establishing the classification only with the clinical information, and the second one with the clinical information and the DE-MRI. In order to avoid any bias between the two tasks, the order of the case is different from the task 1, and moreover, news cases will replace randomly some of them for the classification contest.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Classic proportion of training and test cases for a challenge such as the ACDC challenge [1].

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Unbalanced distribution between normal (1/3) and pathological (2/3) cases, corresponding roughly to real life in managed exams in a MRI department.

### Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

For the DE-MRI, same rules as for task 1.
No specific rules for the annotation of the clinical information, data recorded in a text file for each case with a simple layout.
The request algorithm output will be a binary classification: normal vs pathological case for each exam.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

A specific rule for the myocardial segmentation: the papillary muscles are included in the cavity.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

The different areas are first outlined by an experienced user (a cardiologist with 10 years of experience in cardiology and MRI) and then the segmentation is verified by a second expert (a well-trained biophysicist with 20 years of experience). Same process for the training and test cases.
No specific rules for the clinical information.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

The first expert outlines the contours, and the second goes through every outline and carries out some changes if necessary.
No specific rules for the clinical information.

### Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

To prevent the drawback of the displacement of the heart location between slices due to different breath-holds, the slices are realigned according to the gravity center of the area defined by the epicardial contour. Same process for the training and test cases.
No specific rules for the clinical information.

### Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Noise inside the image could simulate myocardial infarction, and then hamper the classification task.

b) In an analogous manner, describe and quantify other relevant sources of error.

No specific source of error for the clinical information, except that some cases could be ambiguous.

## ASSESSMENT METHODS

### Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

As it is a binary classification, only the classification accuracy is mandatory.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Only one simple metric  to evaluate a binary classification.

### Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

A global ranking according to the classification accuracy. One ranking for each sub-challenge.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Participant teams must provide results for all the cases, otherwise, they will not appear in the final ranking.

c) Justify why the described ranking scheme(s) was/were used.

As it is a classification task in a binary way (normal vs pathological cases), just a global ranking according to the classification accuracy is necessary.

### Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

No missing data handling

No specific statistical methods, except the ranking.

b) Justify why the described statistical method(s) was/were used.

No need to have a deep statistical evaluation.

### Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

No further analysis.

## ADDITIONAL POINTS

### References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

[1] Bernard et al, Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved ? IEEE Trans Med Imaging. 2018 Nov;37(11):2514-2525.
[2] Karim et al, Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images. Med Image Anal. 2016 May;30:95-107.
[3] Kim RJ et al. Relationship of MRI delayed contrast enhancement to irreversible injury, infarct age, and contractile function. Circulation 1999, 100(19):1992 – 2002.
[4] Lalande et al, Evaluation of cardiac structure segmentation in cine magnetic resonance imaging in Multi-modality Cardiac Imaging: Processing and Analysis. Iste, pp. 171–215, 2015.

### Further comments

Further comments from the organizers.

Several members of the organization committee have had some experience in organizing challenges in recent years, and in particular in on-site challenge. We can cite the ACDC challenge (https://www.creatis.insa-lyon.fr/Challenge/acdc/), the IDRID challenge (https://idrid.grand-challenge.org) and the DeepDRiD challenge (https://isbi.deepdr.org).