Multiple Sclerosis Spinal Cord Lesions Detection from MultiSequence MRIs: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Multiple Sclerosis Spinal Cord Lesions Detection from MultiSequence MRIs

Challenge acronym

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

MS-Multi-Spine

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.

Multiple Sclerosis (MS) is a common and potentially debilitating disease affecting around 3 million persons in the world. Currently, Magnetic Resonance Imaging (MRI) plays a central role in this context and in particular allows the identification of MS lesions in the central nervous system. The identification of these lesions on a given MRI image is a complex and mentally demanding task that often leads to an underestimation of disease activity, even for most experienced radiologists. There is thus a need for automated tools that can provide clinicians an aid for accurate and robust identification and quantification of MS lesions. To date, the medical imaging community concentrated its efforts toward the segmentation of the lesions in brain MRI. For this purpose, over the past years, several challenges have been organized to assess the ability of automated methods to detect multiple sclerosis (MS) lesions as compared to manual delineation (The longitudinal lesion challenge; https://smart-stats-tools.org/lesion-challenge, MSSeg: https://portal.fli-iam.irisa.fr/msseg-challenge/, MSSeg2 https://portal.fli-iam.irisa.fr/msseg-2/). These have allowed the community to explore innovative directions. The proposed MS-Multi-Spine challenge aims at offering the possibility to the medical imaging community to extend their methods to spinal cord lesions. This is an innovative challenge both from a clinical and methodological perspective.
1) Clinically, the presence of lesions in the spinal cord has a major prognostic value compared to brain lesions [1]. However, in clinical practice their detection represents a hard task for radiologists. Indeed, MS lesion detection/segmentation in spinal cord MRI is a complex task due to specific characteristics:
- the size of the anatomical structures of interest (around 1cm diameter) resulting in high occurrence of partial volume effects and thus less sharp gradients and contrasts between distinct normal appearing and pathological tissues;
- the occurrence of significant artifacts due to subjects motion and respiration.
As a result, despite its clinical importance, spinal cord MRI is currently under-exploited in patients with MS.

Providing clinicians with tools capable of reliably identifying these spinal cord lesions would therefore be a major added-value.
2) Methodologically, spinal cord lesion detection raises a specific challenge. Indeed, in clinical practice, it is highly recommended to acquire at least two sequences among a set of available sequences, without specific guidelines to date. In practice, depending on the center and context, any combination of existing MR sequences can be provided. In this challenge, that represents a concrete complex case of multisequence datasets, we focus on four commonly used sequences: the sagittal T2 (that is always provided in the challenge and will be considered as the reference to segment), the sagittal STIR, the sagittal PSIR and the 3D MP2RAGE. From a methodological point of view, this is a concrete and paradigmatic case of missing modalities setting where, depending on the case, some modalities may be missing both at inference or training time. To the best of our knowledge, such clinical datasets are still rarely available in medical imaging.

[1] Early imaging predictors of long-term outcomes in relapse-onset multiple sclerosis Brownlee WJ, Altmann DR, Prados F, Miszkiel KA, Eshaghi A et al. Brain, 2019
[2] 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis Wattjes MP, Ciccarelli O, Reich DS, Banwell B, de Stefano N, Enzinger C et al. Lancet Neurol., 2021

Challenge keywords

List the primary keywords that characterize the challenge.

Multiple Sclerosis, Spinal Cord, MRI, Lesion, Instance Segmentation, Detection, Missing Modalities, Missing Data

Year

2025

Novelty of the challenge

Briefly describe the novelty of the challenge.

N/A

Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

N/A

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

None

Duration

How long does the challenge take?

Half a day.

In case you selected half or full day, please explain why you need a long slot for your challenge.

N/A

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 at least 30 challenging pipelines. Our last challenge (MSSeg2, https://portal.fli-iam.irisa.fr/msseg-2/) gathered 30 pipelines from 24 different teams. We expect to attract participants involved in lesions segmentation in medical image, as well as a more general audience interested in the missing modality setting. We will build a new mailing list starting from the MSSeg2 challenge participants (as for now, 142 contacts) plus participants focusing in the missing modalities setting (50 supplemental contacts expected). The first campaign of emails is expected to take place in november 2024.

Publication and future plans

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

For each proposed pipeline, participants will have to submit a technical 4 pages paper describing their approach. The submissions will be reviewed for completeness and clarity. Missing or unclear information will have to be addressed. The resulting proceedings will be made available through an open platform. Similarly to our previous challenges (1,2), we will write a journal paper presenting and discussing the results from the challenge to be submitted to an international journal such as Radiology: AI (https://pubs.rsna.org/journal/ai). The challengers will be asked to participate to a special issue for presenting their approach (see 3).

Beside its original application, the main characteristics of the challenge concerns the use of any combination of existing acquisition sequences at inference as well as at training. Depending on the success of the challenge, in the future we plan to focus our efforts on enriching the data-set by including patients scanned with other sequences and/or other combinations of sequences. To the best of our knowledge, there are only few examples of available multisequence medical imaging dataset involving such a diversity of possible realistic combinations.

1. Commowick O, Kain M, Casey R, Ameli R, Ferré JC, Kerbrat A, Tourdias T, Cervenansky F, et al. (2021), Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. Neuroimage 244:118589.
2. Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, Pop SC, Girard P, et al. (2018), Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Scientific Reports 8:13650.
3. Commowick O, Combès B, Cervenansky F, Dojat M (2023), Automatic methods for multiple sclerosis new lesions detection and segmentation. Front Neurosci 17.

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.

All data will be accessed via the Shanoir platform (https://shanoir.irisa.fr/shanoir-ng/welcome, under the study MS-Multi-Spine). All submitted pipelines will be integrated in the Virtual Imaging Platform (VIP, http://vip.creatis.insa-lyon.fr/), allowing for their execution and evaluation several weeks prior to the challenge day for the announcements of the participant results.

TASK 1: Detect spinal cord lesions

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.

Multiple Sclerosis (MS) is a common and potentially debilitating disease affecting around 3 million persons in the world. Currently, Magnetic Resonance Imaging (MRI) plays a central role in this context and in particular allows the identification of MS lesions in the central nervous system. The identification of these lesions on a given MRI image is a complex and mentally demanding task that often leads to an underestimation of disease activity, even for most experienced radiologists. There is thus a need for automated tools that can provide clinicians an aid for accurate and robust identification and quantification of MS lesions.

To date, the medical imaging community concentrated its efforts toward the segmentation/detection of the lesions in brain MRI. For this purpose, over the past years, several challenges have been organized to assess the ability of automated methods to detect multiple sclerosis (MS) lesions as compared to manual delineation. These have allowed the community to explore innovative directions. The proposed MS-Multi-Spine challenge aims at offering the possibility to the medical imaging community to extend their methods to spinal cord lesions.

This is an innovative challenge both from a clinical and methodological perspective.

1) Clinically, the presence of lesions in the spinal cord has a major prognostic value compared to brain lesions. However, in clinical practice their detection represents a hard task for radiologists. Indeed, MS lesion detection/segmentation in spinal cord MRI is a complex task due to specific characteristics:
- the size of the anatomical structures of interest (around 1cm diameter) resulting in high occurrence of partial volume effects and thus less sharp gradients and contrasts between distinct normal appearing and pathological tissues;
- the occurrence of significant artifacts due to subjects motion and respiration.
As a result, despite its clinical importance, spinal cord MRI is currently under-exploited in patients with MS. Providing clinicians with tools capable of reliably identifying these spinal cord lesions would therefore be a major added-value.

2) Methodologically, spinal cord lesion segmentation raises a specific challenge. Indeed, in clinical practice, it is highly recommended to acquire at least two sequences among a set of available sequences, without specific guidelines to date. In practice, depending on the center and context, any combination of existing MR sequences can be provided. In this challenge, that represents a concrete complex case of multisequence datasets, we focus on four commonly used sequences: the sagittal T2 (that is always provided in the challenge and will be considered as the reference to segment), the sagittal STIR, the sagittal PSIR and the 3D MP2RAGE. From a methodological point of view, this is a concrete and paradigmatic case of missing modalities setting where, depending on the case, some modalities may be missing both at inference or training time. To the best of our knowledge, such clinical datasets are still rarely available in medical imaging.

Keywords

List the primary keywords that characterize the task.

Multiple Sclerosis, Spinal Cord, MRI, Lesion, Instance Segmentation, Missing Modalities, Missing Data

ORGANIZATION

Organizers

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

* Romain Casey: OFSEP (French Multiple Sclerosis Registry) Cohort manager, Univ Lyon, Université Claude Bernard Lyon 1, Hospices Civils de Lyon, Fondation EDMUS, OFSEP, Centre de Recherche en Neurosciences de Lyon, F-69000 Lyon, France
* Benoit Combès: Inria Researcher specialized in MRI processing for MS, Univ Rennes, Inria, CNRS, Inserm; IRISA UMR 6074, Empenn ERL U 1228, F-35000 Rennes, France
* François Cotton: Professor of radiology, Head of imaging group of OFSEP (French Multiple Sclerosis Registry), Past president of the French Society of Neuroradiology/ Affiliations service de radiologie, Hôpital Lyon Sud, Hospices Civils de Lyon, Lyon, France INSA.
* Michel Dojat: Research Director at the Institute National for Health and Medical Research (Inserm) and Deputy scientific director for digital biology and health at Inria, Scientific manager of the FLI-IAM infrastructure Inserm U1216, Université Grenoble Alpes, Inria F-38000 Grenoble France
* Michael Kain: Technical Lead of the Shanoir plateform (medical imaging research platform), Univ Rennes, Inria, CNRS, Inserm; IRISA UMR 6074, Empenn ERL U 1228, F-35000 Rennes, France
* Anne Kerbrat: Associate professor of neurology, specialized in MRI studies for multiple sclerosis. Neurology department, Rennes university hospital, Rennes, France
* Sorina Pop: Head of the VIP platform (web portal for medical imaging applications). INSA Lyon, Universite Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69XXX, LYON, France

b) Provide information on the primary contact person.

benoit.combes@inria.fr
anne.kebrat@chu-rennes.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.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

This challenge has been thought of as a one-time event with a fixed submission deadline. Depending on the challenge reception, the challenge could evolve to a repeated-even, where an incrementally growing number of sequences and sequences combination will be made available.

Challenge venue and platform

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

MICCAI 2025

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

Two plateforms will be used for the challenge:
1) Data sharing with data use agreement (DUA): Shanoir (https://shanoir.irisa.fr/shanoir-ng/welcome).
2) Pipeline submission and evaluation: VIP (http://vip.creatis.insa-lyon.fr/)

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

A webpage has been be created at https://portal.fli-iam.irisa.fr/MS-Multi-Spine/

Participation policies

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

Fully automatic methods only.

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.

In addition to the data provided for the sake of the challenge, only publicly available data and publicly available pre-trained are allowed.

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.

Organizers Institutes may participate in the challenge but will not be listed in leaderboard.

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

No awards and prizes are planned.

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.

All performance results will be publicly announced on challenge day and will be displayed on the challenge website.

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).

For each proposed pipeline, participants will have to submit a technical 4 pages paper describing their approach. The submissions will be reviewed for completeness and clarity. Missing or unclear information will have to be addressed. The resulting proceedings will be made available through an open platform. We will write a journal paper presenting and discussing the results from the challenge to be submitted to an international journal such as Radiology: AI (https://pubs.rsna.org/journal/ai). The challengers will be asked to participate to a special issue of a journal for presenting their approach.

Participant teams that have submitted valid methods and completed the technical paper at the end of challenge will be allowed to use their own performance scores for publication separately after the challenge (12 months embargo time will be imposed after the challenge day).

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.

The participating teams will be requested to submit their algorithms in the form of a Docker container to the Docker repository of the Challenge Pipeline. Detailed instructions for the docker interface and submission will be provided to the participants as we did for previous challenges (https://gitlab.inria.fr/amasson/lesion-segmentation-challenge-miccai21/-/blob/master/SUBMISSION_GUIDELINES.md). In a nutshell, the four following steps will be required :
1. build a Docker or Singularity image containing the method,
2. create a Boutiques (https://boutiques.github.io/) descriptor of the tool,
3. make the image and descriptor available to the VIP team.
4. validate its integration with the VIP team.

Submitted pipelines will be integrated in the Virtual Imaging Platform (VIP, http://vip.creatis.insa-lyon.fr/), allowing for their execution and evaluation several weeks prior the challenge day for the announcements of the participant results.

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.

Only the submitted pipeline will be assessed on the test set. Each team will be allowed to submit a single pipeline.

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

Challenge Website open: July 2, 2024
Dataset annotation : May 2024 - October 2024
Dataset access send to the challenge chairs for review: December 1, 2024
Registration period : December 2024 - May 2025
Training data first release: December 1, 2024
Training data final version release after the challenge chairs review (if needed)/final version: March 1, 2025
Short paper submission deadline: June 15, 2025
Docker submission deadline: June 15, 2025
Announcement of the results: the 23th or 27th of September 2025

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).

All spinal cord MRI data used in this challenge were previously collected as part of research cohorts (OFSEP, MAPMS, MSTRACTS). All patients with MS included in these cohorts had given written informed consent to the use of their medical imaging data for research purposes (OFSEP: NCT02889965, MAPMS: NCT04918225, MSTRACTS: NCT04220814).
Each participating team will be required to accept data usage agreement (DUA) using the shanoir web plateform before being authorized to access the training dataset. The planned DUA will be similar to those used to share OFSEP data in a similar context (see section DUA).

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)

Below is the planned DUA, participants will have to accept and comply with in order to access the data:

Access to OFSEP and CHU de Rennes MRI used for the MICCAI 2025 challenge

I request access to the data collected in the digital repository of the OFSEP/CHU de Rennes MICCAI 2025 Cohort provided in the context of the MICCAI 2025 challenge.
By accepting this agreement, I become the data controller (as defined under the European GDPR) of the data that I have access to, and I am responsible that I access these data under the GDPR obligations and the specific following terms:
1. I will comply with all relevant rules and regulations imposed by my institution and my government. This agreement never has prevalence over existing general data protection regulations that are applicable in my country.
2. I will not attempt to establish or retrieve the identity of the study participants. I will not link these data to any other database in a way that could provide identifying information. I shall not request the pseudonymisation key that would link these data to an individual's personal information, nor will I accept any additional information about individual participants under this Data Use Agreement.
3. I will not redistribute these data or share access to these data with others, unless they have independently applied and been granted access to these data, i.e., signed this Data Use Agreement. This includes individuals in my institution.
4. When sharing secondary or derivative data (e.g. group statistical maps, learnt models or templates), I will only do so if they are on a group level, and information from individual participants cannot be deduced.
5. I will reference the specific source of the accessed data when publicly presenting any results or algorithms that benefited from their use: (a) Papers, book chapters, books, posters, oral presentations, and all other presentations of results derived from the data should acknowledge the origin of the data as indicated in the Terms of use below (b) Authors of publications or presentations using the data should cite relevant publications describing the methods developed and used as described in the Terms of use below (c) Neither the [Research centre/University Department] or [University] or [Institution], nor the researchers that provide this data will be liable for any results and/or derived data.
6. I will register my [Research centre/University Department] or [University] or [Institution] in the agreement form and I accept that it is cited on the OFSEP website; I will register using my professional email address.
7. I will have the right to use this dataset for a period of 3 years starting from the date when access to the dataset will be granted. If I need to use this data for more time, I will have to ask for an extension on this website (https://shanoir.irisa.fr). Otherwise, I will have to delete them from my disk after this period of three years.
8. I will not publicly present any results or algorithms that benefited from the use of these data until 12 months after the challenge results have been presented at the MICCAI conference.
9. I will not use this data or a derivative product of it for a commercial use – if looking for a commercial use of the dataset or unsure, please contact OFSEP (projects@ofsep.org).
10. I will inform OFSEP and CHU de Rennes of the publication of my article with its references via the email addresses publications@ofsep.org.
11. I agree to be contacted from time to time by OFSEP and CHU de Rennes staffs in charge of projects and publications in order to follow the progress of my work.
12. Failure to abide by these guidelines will result in termination of my privileges to access these data.

Terms of use
When using part or all of this data, please adhere to the following guidelines:

1) Indicate in Methods

Data were generated i) by participating neurologists in the framework of Observatoire Français de la Sclérose en Plaques (OFSEP), the French MS registry (Vukusic et al. 2020). They collect clinical data prospectively in the European Database for Multiple Sclerosis (EDMUS) software (Confavreux et al. 1992). MRI of patients were provided as part of a care protocol. Nominative data are deleted from MRI before transfer and storage on the Shanoir platform (Sharing NeurOImagingResources, shanoir.org).
Vukusic S, Casey R, Rollot F, Brochet B, Pelletier J, Laplaud D-A, et al. Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France. Mult Scler. 2020;26(1):118–22.
Confavreux C, Compston DAS, Hommes OR, McDonald WI, Thompson AJ. EDMUS, a European database for multiple sclerosis. J Neurol Neurosurg Psychiatry 1992; 55: 671-676.
ii) by data collected as part of MAPMS and MSTRACTS research protocols (MAPMS: NCT04918225, MSTRACTS: NCT04220814). Nominative data are deleted from MRI before transfer and storage on the Shanoir platform (Sharing NeurOImagingResources, shanoir.org).

Please cite the challenge dataset description article, the challenge paper, and the MAPMS and MS TRACTS cohort description in any publication using a part or all of the dataset images, when they will become available (see https://portal.fli-iam.irisa.fr/ms-multi-spine/ for updates).

2) Add to Acknowledgments
This work was carried out i) in collaboration with The Observatoire Français de la Sclérose en Plaques (OFSEP), who is supported by a grant provided by the French State and handled by the Agence Nationale de la Recherche, within the framework of the France 2030 program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation; ii) in collaboration with the CHU de Rennes, the promoter of the MAPMS and MS TRACTS cohorts. These cohorts were funded by the ARSEP foundation and by CHU de Rennes CORECT grants.

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.

The code to compute the performance metrics and to rank the methods will be released in November 2024.

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

Participants will be required to submit their methods to the VIP platform as Docker containers. Participating teams code will not be required to be publicly accessible.

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.

We do not report any conflict of interest.

No sponsoring/funding is provided for this challenge.

Only the members of the organizing committee and associated technical staff that will need to manipulate the annotations will have access to the final annotations on the test set. This will happen in two phases: first to build the ground truth masks (thus prior to dataset release) and finally to compute methods performances (thus after model evaluation).

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, Prognosis, Follow-up, Monitoring, Treatment response Evaluation, Research

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

Instance Segmentation, Detection, Localization

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 target cohort consists of patients with multiple sclerosis with different ages, disease durations, phenotypes and lesion involvements.

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

The challenge cohort consists of patients with multiple sclerosis with different disease durations, phenotypes and lesion involvements whose spinal cord MR imaging consists of at least two sequences, including a t2 sagittal scan and acquired on a 1.5 or 3T MRI scanner.

Imaging modality(ies)

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

MRI scans acquired after 2015 to reflect current clinical practice, using various 1.5T and 3T MRI scanners from different manufacturers, with at least 2 spinal cord sequences acquired to match the current international recommendations, including at least sagittal T2 sequence (most commonly used) and at least one of sag-STIR, sag-PSIR or 3D-MP2RAGE.

Context information

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

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

Sequence type (among the 5 possible ones), position of the slab (top or bottom), MRI scanner brand (among the 3 possible ones).

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

None

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.

Spinal cord MRI acquisitions from different sequences and with different fields of view. Depending on the acquisition, a given slab can span a variety of range in the Head-Foot axis, with various anatomical coverage (possibly starting in patients brain and ending lower than the last lumbar vertebra).

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 target consists of all T2 hyperintense lesions located anywhere in the spinal cord, starting from the top of the C1 cervical level to the last lumbar level. In the challenge cohort, the target lesions are restricted to those contained in the acquired volume slab (never encompassing the whole spinal cord for a given acquisition).

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)).

The evaluation has been designed to assess method's ability to detect MS spinal cord lesions. The objective of the method is to localize/segment each spinal cord MS lesion and to assign to each detected instance a probability of being of a lesion. The methods' performances will be assessed using the mean sensitivity averaged among the five false positive rates 0.25, 0.5, 1, 2 and 3. The achieved sensitivity for each of the five levels will be estimated as part of the evaluation procedure individually for each pipeline from the probabilities assigned to each predicted instance. The localization criterion used to match ground-truth and predicted instances will be mask-IoU with a relatively low threshold (0.2).

A given method will be asked to output a mask of label as well as a corresponding csv file assigning a probability to each lesion's label (one row per lesions).

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).

MR images were acquired from multiple 1.5 and 3T MRI scanners (Siemens, Philips and GE) from different centers. Overall, data originates from 21 different scanner models distributed among 35 different clinical centers.

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

To cover a large portion of the spinal cord in a reasonable acquisition time , most spinal cord sequences are acquired in the sagittal plane with typical spacing of 3mm and in-plane (sagittal) resolution of 0.6x0.6mm. In the context of this challenge, we will use three different sagittal sequences, T2 weighted (T2), Phase-Sensitive Inversion Recovery (PSIR) and the Short Time-Inversion Recovery (STIR). Depending on the center/scanner we can find a variety of resolution settings. Below is displayed the mean and range resolutions in the challenge data set for each sequence:

- Mean T2 resolution: 3 (min: 2.2, max 3.9) x 0.5 (min:0.3, max: 0.9) x 0.5 (min:0.3, max: 0.9) mm3
- Mean T2 STIR resolution: 2.2 (min: 0.9, max 3.3) x 0.6 (min:0.3, max:0.8) x 0.6 (min:0.3, max: 0.8) mm3
- Mean T2 PSIR resolution: 3.2 (min: 2.5, max 4.4) x 0.5 (min:0.3, max: 1.2) x 0.5 (min:0.3, max: 1.2) mm3

In addition to these common sagittal sequences, we used a fourth sequence, the Magnetization Prepared 2 Rapid Acquisition Gradient Echoes (MP2RAGE). While only the spinal cord will be used in this challenge, these sequences offer the possibility to image both the brain and the top of the spinal cord in a reasonable time (8 mins) with nearly isotropic voxels (around 1 mm3).

In addition to differences of resolutions, these sequences overall offer different characteristics of intensity contrasts and dynamics and propensity to artifacts.

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.

Data were acquired in several French clinical centers over the past 10 years. Overall, they come from 35 different clinical centers.

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).

All these images were acquired either for patient-care and/or clinical research purpose and were thus acquired by professional MRI technicians.

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.

A case refers to a given set of acquisitions (consisting at least of a sagittal T2 acquisition and another acquisition) imaging a same anatomical location from a given patient. It must be noted that, for a given case, the spatial coverage can be slightly different from one acquisition to another. The spatial coverage of the sagittal T2 will always be the one considered for annotating the lesions and evaluating the methods. Cases from the training and testing set are segmented using the same methodology.

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

The 200 set of acquisitions (each with two or three available sequences) will be split into 100 (50%) training case and 100 testing cases.

# Repartition of sequences configuration in training and testing set

##Training (overall 100)

(t2, stir) : 50 cases
(t2, psir) : 25 cases
(t2, mp2rage) : 25 cases
(t2, stir, mp2rage) : 0 case

##Testing (overall 100)

(t2, stir) : 40 cases
(t2, psir) : 20 cases
(t2, mp2rage) : 20 cases
(t2, stir, mp2rage) : 20 cases

# Repartition of scanner brands in training and testing set

##Training (overall 100)

Siemens : 67 cases
Philipps: 33 cases
GE: 0 case

##Testing (overall 100)

Siemens : 64 cases
Philipps: 20 cases
GE: 16 cases

# Repartition of scanner field strenght

##Training (overall 100)

1.5T : 45 cases
3T: 55 cases

##Testing (overall 100)

1.5T : 42 cases
3T: 58 cases

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

In our opinion, a training set of 100 cases for training and 100 cases for testing is a reasonable sample size for this challenge.

First, generally speaking, this number is superior to what we observed in most challenges as well as publications involving MS brain lesions (Isbi-2015: 82 sessions overall from 19 unique patients (4 patients in train/15 in test), MSSeg-2016: 45 sessions overall (20 train/25 test) from 45 unique patients, MSSeg2-2021: 100 sessions overall (40/60) from 100 unique patients, MSLegSeg-2024 (to come): overall 150 MRI sessions from 90 independents patients). In particular, a testing set of 100 cases is reasonable to assess the performance of the models on a variety of conditions. It must be noted that at this step we don't know the number of lesions/lesion-voxels included in our data set, but based on our experiences on MS spinal cord, we extrapolate a total lesions number of 250 lesions with total volume 37500 mm3 in the training set (and similarly for the testing set).

Second, SC imaging is less common and their annotations require more expertise than for the brain. Thus, while 100 cases may be not sufficient to develop models close to optimality, we think relevant to consider that working with a limited sample size is part of the setting. Similarly, it should be considered that while all of 100 training cases include a T2 acquisition, the three other sequences are less represented in the training data set. Again, we consider that this relative low number of additional sequences is one characteristic of the proposed challenge.

Finally, it must be outlined that to the best of our knowledge, this is the very first effort toward providing of a SC MS lesions dataset to the community. Providing 200 high quality annotated cases, each annotation involving 5 experts, consists of a substantial effort to bring this data to the community.

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.

First, to assess the ability of methods to deal with unknown combinations of sequences, the combination of sequences (T2, STIR, MP2RAGE) will be included only in the test set.
Then, to assess the ability of methods to generalize out of the training set, all scans from GE scanners will be included only in the testing set.
The training and testing set displays a balanced repartition of 1.5T and 3T scanners.

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

N/A

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.

To ensure high quality annotations, each case will be independently annotated by four different annotators (radiologists with specialization in neuroradiology and at least 2 years' experience). Each annotator will perform a voxel-wise delineation of the spinal cord MS lesions (the top of the C1 vertebrae being considered as an upper bound for the spinal cord) on the sagittal T2 images with the help of all other sequences available. The ground-truth will be finally built by an external highly experienced expert (>10 years of experience on reading SC MRI of MS patients) that will accept or reject each candidate lesions raised by at least one of the 4 annotators. The final lesion delineation of each accepted lesion will be obtained by merging the different segmentation masks available using a max-voting operation.

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.

Each of the 4 annotators will benefit from
1) an online tutorial with a precise labeling protocol including:
a. details of the procedure for using itksnap (how to open the different images, adjust contrast, which brush and size to choose for labeling lesions, how to save the lesion mask)
b. examples of good/bad MS lesions delineations on the sagittal T2.
The annotation protocol can be accessed via this link :
https://docs.google.com/document/d/1GeiiwQEJIjRjZNL-sagNU3burb5T1hW3a1PXczR1TcE/edit?usp=sharing

2) A face-to-face training to show live examples of spinal cord lesion delineations with itksnap and answer questions from annotators

The first 4 annotations provided by annotators for this proposal will be reviewed by the external highly experienced expert, and individual feedback will be given to each expert.

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.

Each case will be segmented by four medical doctors (radiologists with specialization in neuroradiology and at least 2 years' experience) using the itksnap software (itksnap.org/pmwiki/pmwiki.php).

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 segmentation from the 4 experts will be fused by a two-steps process: first, each candidate lesions raised by at least one of the 4 annotators will be accepted or declined by an external highly trained expert (>10 years of experience on reading SC MRI of MS patients). In case of doubt, acquisitions performed before or after the considered time point in the patient follow-up will be used by the external expert to accept or reject a given lesion. Second, for each accepted lesion, the final voxel-wise delineation will be produced by majority voting.

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.

In addition to the raw SC images, two versions of the data will be provided to the challengers:
- Preprocessed images: in a nutshell, for a given case, all raw images will be sequentially i) reoriented in a sagittal orientation, ii) resampled in a single frame associated to a fine resolution (0.5mm3) and iii) zeroed outside a square area of side 35 mm centered (slice-wise) on the spinal cord barycenter.

- Registered and preprocessed images: the preprocessed images after having applied a rigid followed by a highly regularized non-linear registration will also be provided.

The associated scripts and some intermediate results will be provided to the challengers. The resulting images can then be used directly for training or further processed. We emphasize that there are a number of choices in the preprocessing/registration that are likely to have an impact on model/learning performances and challengers will be allowed to develop their own preprocessing to be applied on the raw and/or preprocessed images. When a specific preprocessed stage is introduced, the authors need to indicate and document it in the associated method description.

For the evaluation stage, the participant's docker will be interfaced to the relevant data (raw images and/or preprocessed images and/or preprocessed and registered) though different command line arguments.

Finally, if necessary, for each acquisition a given amount of top slices of the MRI will be removed to exclude the patients head/face from the acquisition field of view.

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.

Sources of Errors for Lesion detection:

MS lesion identification and segmentation in spinal cord MRI is a complex task mainly due to low contrast between normal appearing tissue and lesions and a high occurrence of artifacts. In practice, it is very common to observe doubtful intensities in a given acquisition without being able to attribute its to real tissue abnormality or to artifacts or partial volume effects.

In all the settings studied in this challenge, at least two sequences will be used to annotate lesions, which will allow to limit this concern. However even in this case, the inter annotator error is high at the lesion scale (i.e. detected or not detected). We estimated an inter-rater Lights kappa of 0.55 for lesion detection in a study under progress. The use of 4 experts for each case followed by the external expert adjudication with access to external (follow-up) data will help to mitigate these errors. It will also avoid the risk of incidental annotations (e.g. misclick) that may arise in the context of large scale data annotations.

Sources of Errors for Lesion Contouring:

For a given lesion, defining its outlines is also a complex task due to partial volume effect as well as the practical difficulty to adjust contrasts due to lack of sharp gradient between normal and pathological tissues. Such annotation errors are inherent to the data and can be dealt with by the proposed methods. For a given annotated lesions, the inter-rater contouring variability will be assessed from the annotated data.

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

Patients BMI and positioning into the MRI scanners may influence the quality of the acquisitions and of the resulting annotations ; especially for lower sagittal coverage.

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 evaluation has been designed to assess methods' ability to detect/localize lesions. Methodologically, the setting is those of an instance segmentation problem where probabilities are assigned to inferred instances and where ground-truths are provided as segmentation masks.

The localization criterion used to match ground-truth and predicted instances will be mask-IoU with a relatively low threshold (0.2). A too high value (e.g. 0.5) would imply a good overlap between estimated and ground-truth maps for a lesion to be considered as detected, whereas we know that such outlines are of little interest and difficult to assess precisely (which is of exacerbated by the fact that lesions can be of small size). By contrast, a too low value would lead to considering lesions detected even for cases where the outlines -while overlapping- have only little to do with the hyper-intensity of interest. The dependency between this value and the resulting model performances and rankings will be explored in supplemental analysis.

To assess the performance of lesion-wise detection, we will use a FROC-based metric consisting in computing the mean sensitivity averaged among the five false positive rates 0.25, 0.5, 1, 2 and 3. The achieved mean sensitivity for each of the five levels will be estimated as part of the evaluation procedure individually for each pipeline based on the probabilities assigned to each inferred lesion.

In cases where multiple predictions match a single ground truth instance, one will count as a TP and the extra predictions matching the same GT instance won't be penalized as FP.

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

First, we focus on a detection-based metric, thus aiming at measuring the ability of methods to detect/localize lesions without the need to precisely estimate their outlines. In clinical practice, the precise delineation of lesions is rarely of interest and in many cases consists of an ill-posed problem due to partial volume effect. By contrast, being able to robustly detect each lesion is closer to the clinical objective of MRI examination.

Then, the selected detection-based metric is relevant for several reasons. First, it matches the practical objective of interest (i.e. achieving optimal sensitivity while restraining a practically reasonable number of false positives). Second, as a multi-threshold metric, it allows a characterization of the model performances over a range of reasonable settings and potentially increases the robustness of the pipelines comparison. Finally, its value (i.e. a sensitivity level) has a direct interpretation of primary usage.

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.

The methods will be directly ranked according to their averaged mean sensitivity score.

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

Data on which a pipeline does not output any valid files will be associated to 0 sensibility for the given case.

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

For each FP rate, sensitivity rates are averaged on all cases disregarding their number of lesions. Thus, a false detection or a missed lesion on a case with a single lesion has a much bigger impact on the overall score than a on a case with many lesions. This choice is in line with the clinical setting, where it is much more important to find all lesions on patients with few lesions than on patients with many.

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.

In addition to the ranking described above, the models and experts will be analyzed through different aspects:

1. Performance comparison with respect to experts

Once we got the final results, we will assess differences of performances between the four experts and the different methods. This will be performed in two ways:
- First, by selecting each method decision threshold so that the subsequent FP rate will match those from the expert with the highest FP rate. Then the associated sensibility will be compared between each method and the expert. In particular, for each method, we will compute the p-value associated with the null hypothesis "no difference in sensitivity with respect to the expert" using pairwise Wilcoxon tests. The same analysis will be performed again using the expert with the lowest FP rate.
- Second, using the F1 score from the experts and the methods, where for each method the decision threshold will be computed so that the F1 score is maximal. Then, for each expert and method, we will compute the p-value associated with the null hypothesis "no difference in mean F1 with respect to the performance of the best expert" using pairwise Wilcoxon tests.

This comparison will be performed by using all cases as well for each of subgroups of sequence settings (i.e. t2+stir, t2+psir, t2+mp2rage, t2+mp2rage+stir).

2. Added-value of multiple sequences on model performance

For each case, we will assess model performances by using a cumulative number of sequences (i.e. for (T2, STIR, MP2RAGE) we will test models on (T2, STIR), (T2, MP2RAGE) and (T2, STIR, MP2RAGE)) and test for difference of performances when using the different combinations using pairwise Wilcoxon tests (leading to 3 pairwise comparisons). To avoid a too large number of comparisons/tests performed, this analysis will be limited to the top 4 performing methods.

3. Effect of patient, scanner and lesions characteristics on performances

Multiple parameters can affect the performances of participant methods. In particular, images have varying qualities, the number of lesions can vary significantly between cases, the size and location of lesions and the number and type of available sequence is likely to affect their visibility. Different covariates will be included: i) a level cofactor associated to out-of-training scanners (i.e. GE scanner) will be considered to assess a potential difference of performances between cases from scanners used at training and scanners (GE) not in the training set, ii) a level cofactor associated to the method making the detection, iii) the volume of the lesion considered, iv) the overall lesion volume in the case of interest and v) the number of available sequences for the case considered. To investigate the association between the propensity of methods to make false positives or false negatives and these different factors, we will fit logistic models. To optimize the informativeness of this exploration, this analysis will be performed both for the detections obtained for the highest (4) and lowest (0.25) explored false positive rates. To avoid a to large number of comparisons/tests performed, this analysis will be limited to the top 4 performing methods.

4. Other performance metrics

We will also describe performance of methods for other metrics than the one used for ranking. In particular, we will explore performances in terms of best F1 score (estimated at optimal working point as part of the evaluation procedure), lesion-wise sensitivity for each of the investigated FP levels and, for methods building segmentation masks, DICE index and estimated lesion volume. Finally, we will also explore the ability of methods and experts to classify correctly cases in one of the three following categories: i) no new lesion, ii) one or two new lesions, and iii) three or more new lesions (for the different FP levels).

All these analyses will be performed using the R software. All tested hypotheses will be explanatory and performed in a Fisherian framework and no adjustment for multiple comparisons will be performed.

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

Wilcoxon test: Given that – strictly speaking- our data are not continuous (e.g. sensitivity can only be a fraction of the number of lesions present in the considered case) we will use a Wilcoxon test to test statistical dominance between groups instead of a student test to test for equality of mean between groups.

Logistic regression: we will be interested in studying the outcomes true lesion detected to study propensity of methods to make false negative, and false lesion detected, to study propensity of methods to make false negative. A logistic regression model with fixed effects (for pairing to cases and other cofactors) is the appropriate model for such a setting.

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.

The notion of detection between instances is based on the definition of a threshold for the minimal IoU. In this challenge, we set this value to 0.2. To contextualize our results, we will investigate the robustness of the estimated performances of subsequent methods ranking when the threshold deviate for the chosen value.

Similarly, we will investigate the robustness of estimated performance and ranking depending on whether or not we chose to penalize as false positive extra matching predictions when more than one predicted instance matche a single GT instance (no penalization will be applied in the baseline performance evaluation).

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.

N/A

Further comments

Further comments from the organizers.

N/A