Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure

Olivier Commowick, Christian Barillot

MSSEG Miccai 2016 Challenge:
Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure

Olivier Commowick, Christian Barillot and FLI / OFSEP

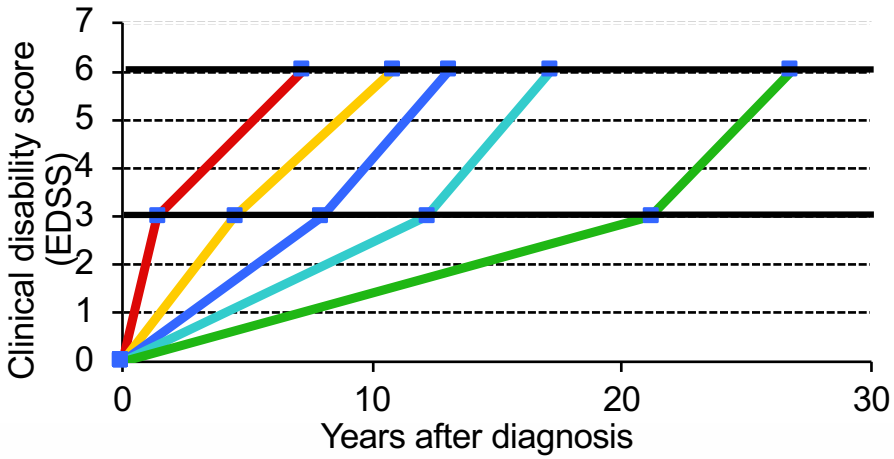
Workshop FLI-SFR – October 11, 2018

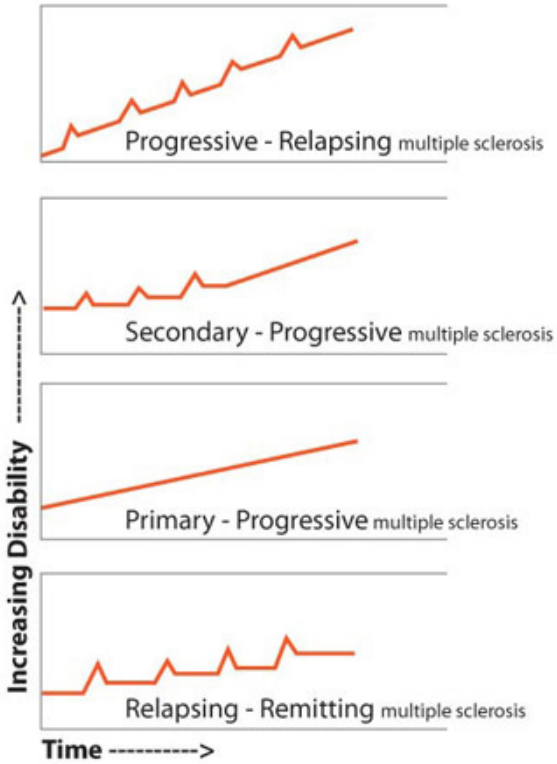
***Commowick O**, Istace A, Kain M, Laurent B, Leray F, Simon M, Pop SC, Girard P, Améli R, Ferré JC, Kerbrat A, Tourdias T, Cervenansky F, Glatard T, Beaumont J, Doyle S, Forbes F, Knight J, Khademi A, Mahbod A, Wang C, McKinley R, Wagner F, Muschelli J, Sweeney E, Roura E, Lladó X, Santos MM, Santos WP, Silva-Filho AG, Tomas-Fernandez X, Urien H, Bloch I, Valverde S, Cabezas M, Vera-Olmos FJ, Malpica N, Guttmann C, Vukusic S, Edan G, Dojat M, Styner M, Warfield SK, Cotton F, **Barillot C**.*

Nature Scientific Report; 8(1):13650, 2018. doi: 10.1038/s41598-018-31911-7.

Background: multiple sclerosis

- Highly variable evolution
  - Clinical classification in 4 types
  - Two main stages
    - Early: variable evolution
    - Later: parallel evolution





Leray, E. et al., 2010. Evidence for a two-stage disability progression in multiple sclerosis. Brain, 133 (7), 1900 - 1913.

Lesion segmentation in MS

- Lesion load and lesion count crucial in MS
  - Part of diagnosis (McDonald criteria)
  - Evaluation of drug effectiveness

- Delineation of lesion tedious
  - Manual → time consuming
  - Subject to intra- / inter-individual variability

➔ Automatic segmentation is key

Thompson, A. et al., 2017. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet, 17 (2), 162 - 173.

Why a segmentation challenge?

- A huge number of automatic segmentation methods
  - Tissue classification & outlier detection
  - Machine learning (random forests, deep, etc.)
  - Many others
- Large variety of modalities used
  - T1, T2, FLAIR, PD…
- Large variety of implementations
  - GPU, Matlab, Python, C++ …

5 surveys in the last 5 years involving 50+ methods

Why a segmentation challenge?

- Evaluation complicated
  - Each method evaluated on a specific set
  - No comparison possible
- The challenge concept
  - Have all methods evaluated on a common dataset
  - Examples: MICCAI 2008, IEEE-ISBI 2015
  - Main drawbacks
    - Possibility to adapt parameters to each patient
    - Ground truth not well defined

Styner et al., 2008. 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. Insight journal.
Carass et al., 2017. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. Neuroimage, 148, 77-102.

An OFSEP and FLI challenge @ MICCAI

- Evaluation objectives
  - Evaluate algorithms developed in the community
  - In a well defined computational framework (FLI)
    - Same set of parameters for all images
    - With respect to a solid ground truth
- Additional objectives (OFSEP)
  - Evaluate lesion segmentation algorithms for MS
  - Fully automatic, on standardized images
    - Standardized but different centers

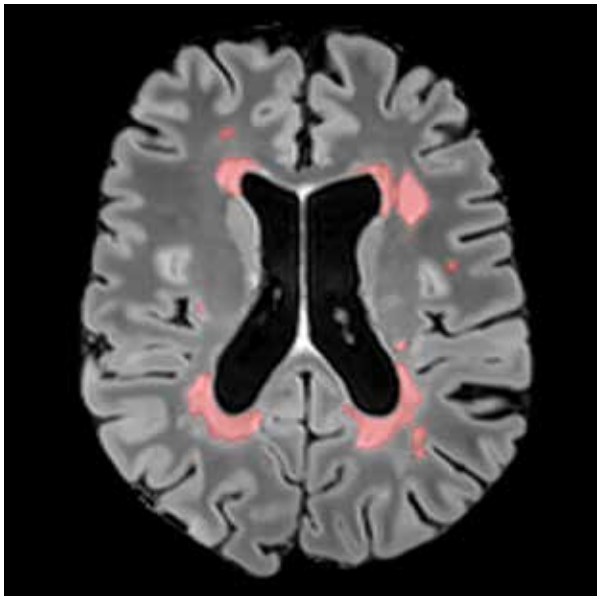
http://www.ofsep.org
Cotton, F., Kremer, S., Hannoun, S., Vukusic, S., Dousset, V., 2015. OFSEP, a nation-wide cohort of people with multiple sclerosis: Consensus minimal MRI protocol. Journal of Neuroradiology 42 (3), 133 – 140.

MICCAI challenge: The Data

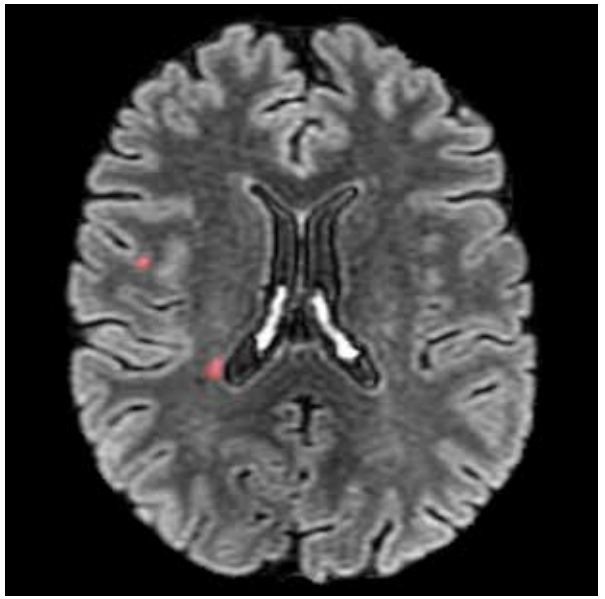
- Challenge data
  - 53 patients from 4 different scanners
  - Modalities: 3DFLAIR, T2/DP, 3DT1, 3DT1-Gado
    - ➢ *OFSEP consensus*
  - **7 manual segmentations for each patient**
- Two datasets drawn
  - Training (open): challengers tune their algorithms
  - Testing (closed): evaluation database

Center / #exams	Training set	Testing set
01 - Siemens Verio 3T (Rennes)	5	10
03 - GE Discovery 3T (Bordeaux)	0	8
07 - Siemens Aera 1.5T (Lyon)	5	10
08 - Philips Ingenia 3T (Lyon)	5	10
Total	**15**	**38**

Dataset examples (*with experts consensus*)



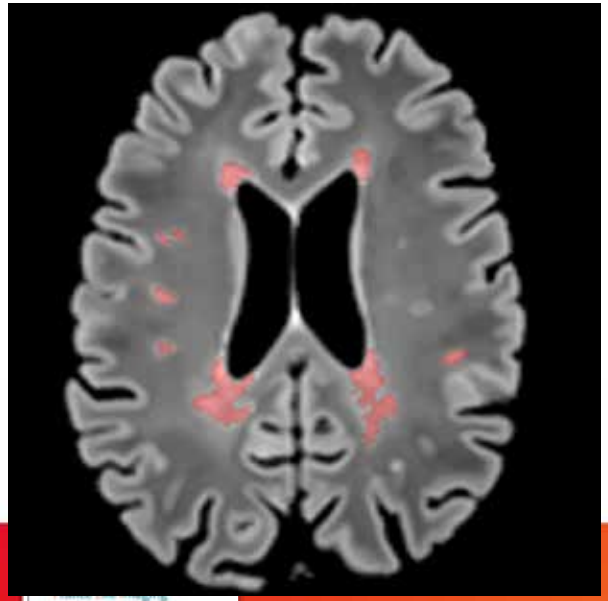
FLAIR from center 01



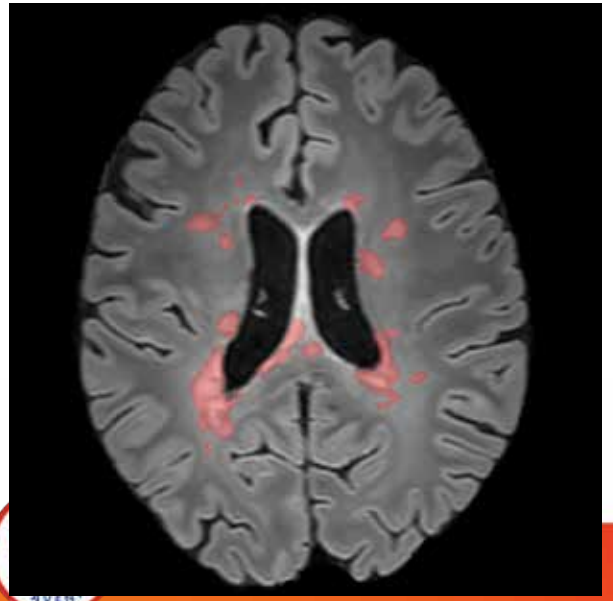
FLAIR from center 03

↓

Not in the Training



FLAIR from center 07



FLAIR from center 08

OFSEP
Observatoire Français de la Sclérose en Plaques

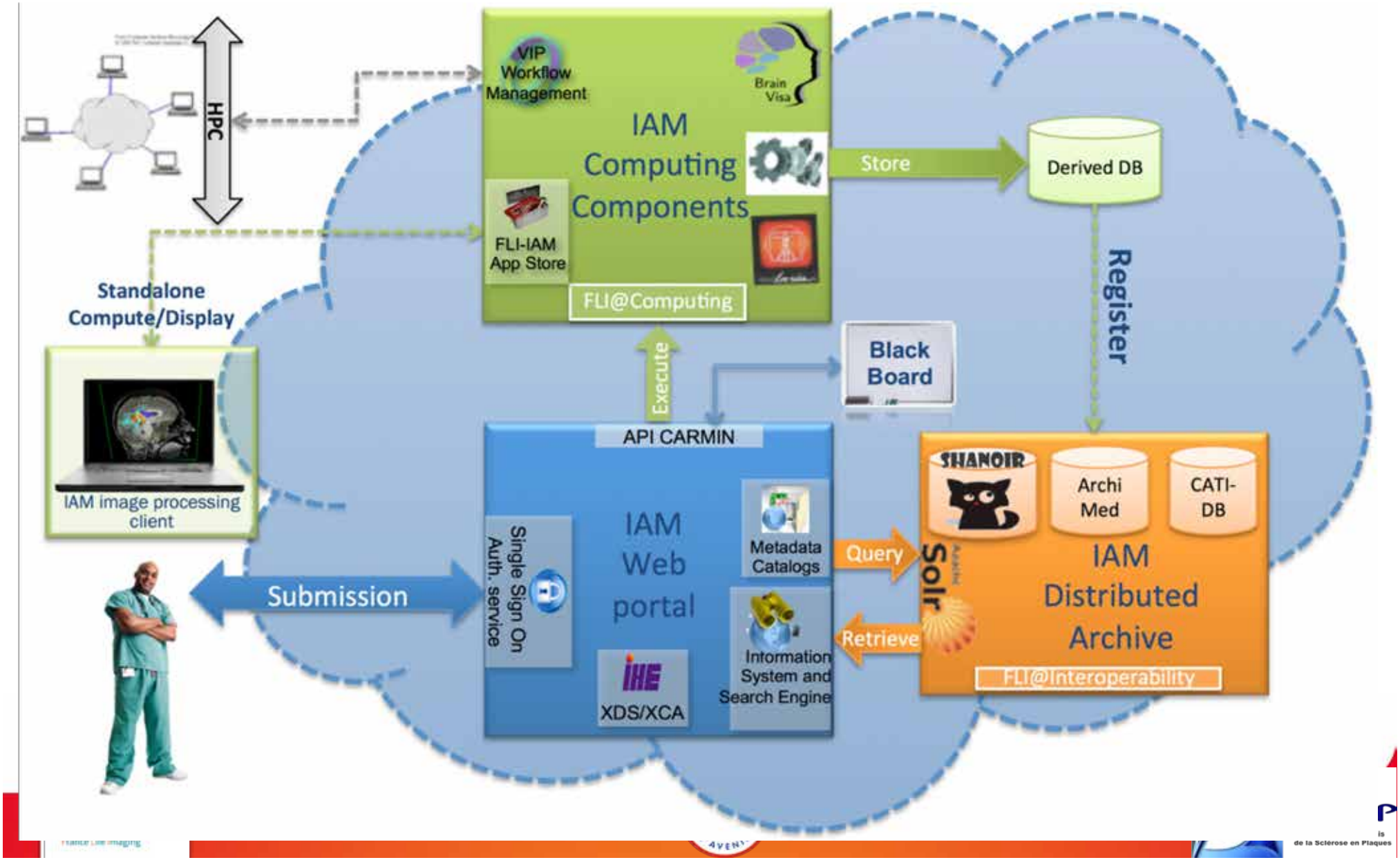
A well defined execution and evaluation framework

- Pipelines provided by the challengers
  - Black box (docker) including their optimal parameters
  - Parameters chosen or optimized on training set

- Pipelines started automatically on testing set
  - On France Life Imaging (FLI-IAM) computing platform
  - By FLI-IAM project engineers
  - Ensures a uniform set of parameters on the whole testing database

https://portal.fli-iam.irisa.fr/msseg-challenge/overview

France Life Imaging computing platform



Challenge participations

- Thirteen pipelines including a variety of algorithms
  - Machine learning:
    - Random forests
    - Deep learning
  - Model Inference (Bayes, Markov, …):
    - Tissue classification approaches

- Training phase: 2 months (*at home*)
- Integration phase: 3 to 4 months (*on FLI-IAM system*)
  - Docker packaging and integration help by FLI
- Evaluation (independent from challengers): 2 months

Which evaluation? Metric categories

- Evaluation of MS lesions segmentation: tough topic
  - Which ground truth? → LOP STAPLE consensus
  - What is of interest to the clinician?

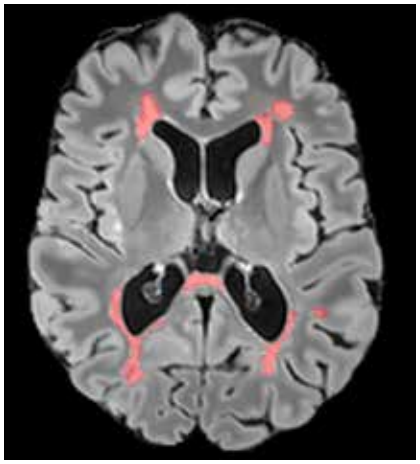
- Two metric categories:
  - Detection: are the lesions detected, independently of the precision of their contours? ➔ *F1 score*
  - Segmentation: are the lesions contours exact?
    - Overlap ➔ *Dice score*
    - Surface-based measures ➔ *Mean surface distance*

A. Akhondi-Asl et al. A Logarithmic Opinion Pool Based STAPLE Algorithm for the Fusion of Segmentations With Associated Reliability Weights. IEEE TMI, 33(10):1997–2009, Oct 2014.
https://portal.fli-iam.irisa.fr/msseg-challenge/evaluation

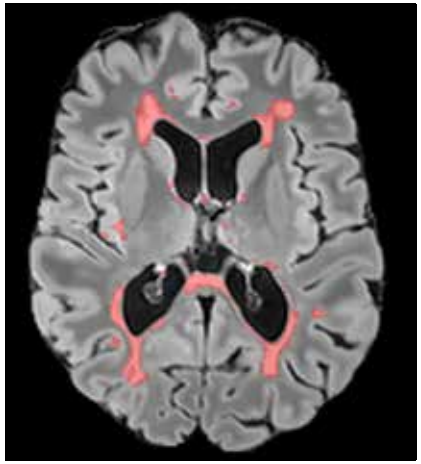
No lesion case results

Evaluated method	Lesion volume (cm^3)	Number of lesions
Team 1	8.25	18
Team 2	**0**	**0**
Team 3	**0**	**0**
Team 4	N/A	N/A
Team 5	28.44	522
Team 6	0.47	7
Team 7	5.99	168
Team 8	**0**	**0**
Team 9	2.55	33
Team 10	11.09	31
Team 11	3.44	42
Team 12	0.06	1
Team 13	0.07	4

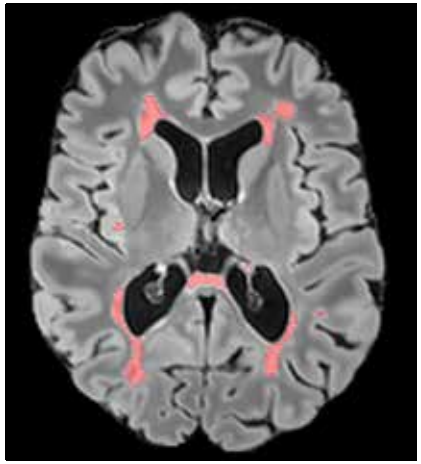
Visual results for center 01



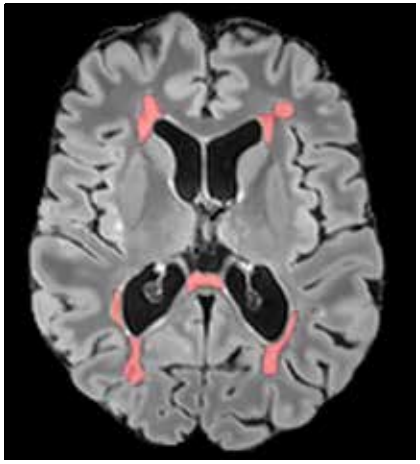
Consensus



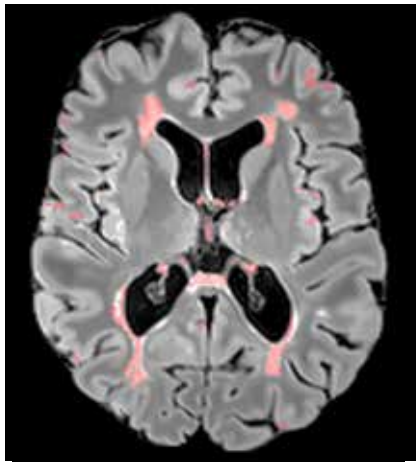
Team 7



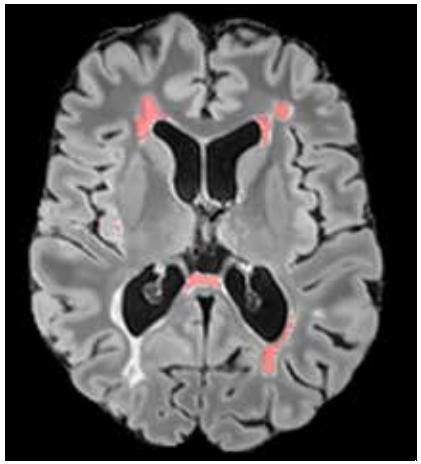
Team 8



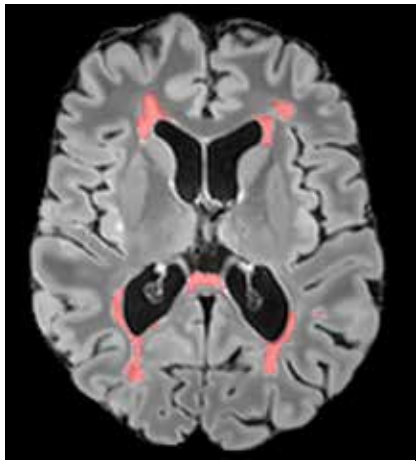
Team 9



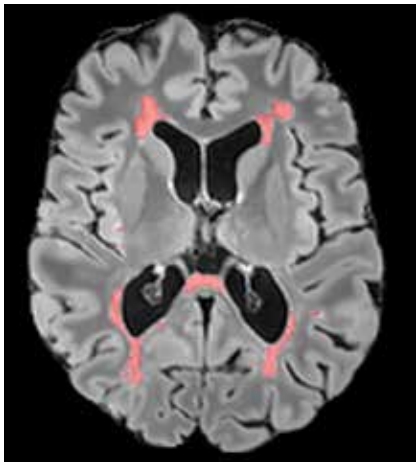
Team 10



Team 11

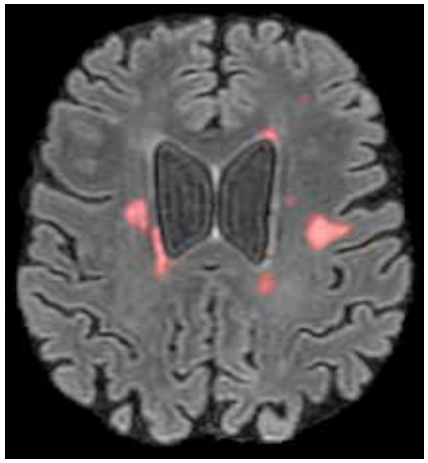


Team 12

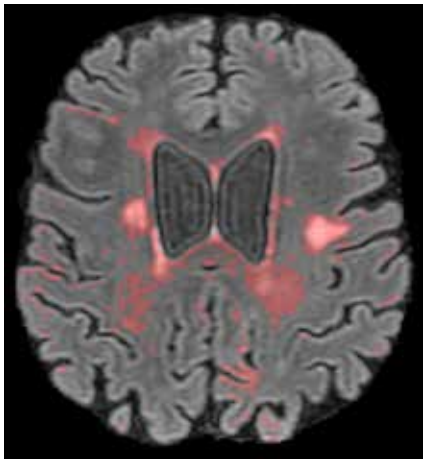


Team 13

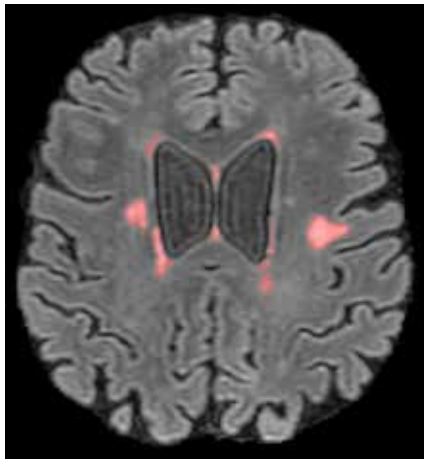
Visual results for center 03 (*not in the training phase*)



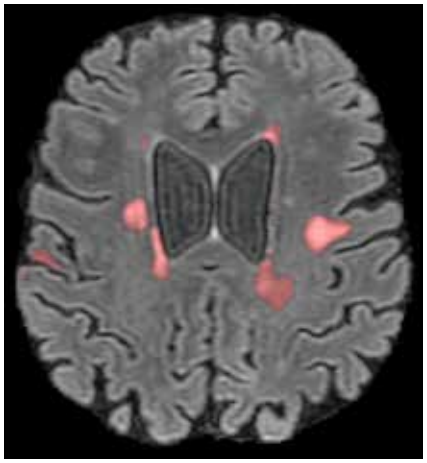
Consensus



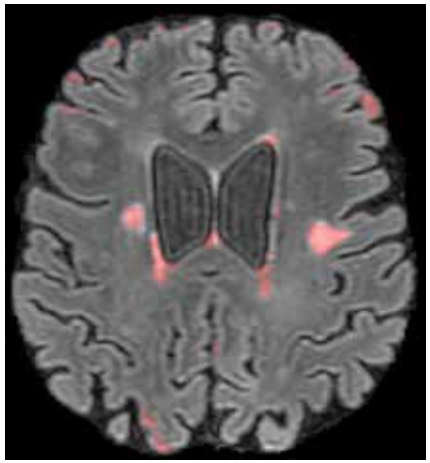
Team 7



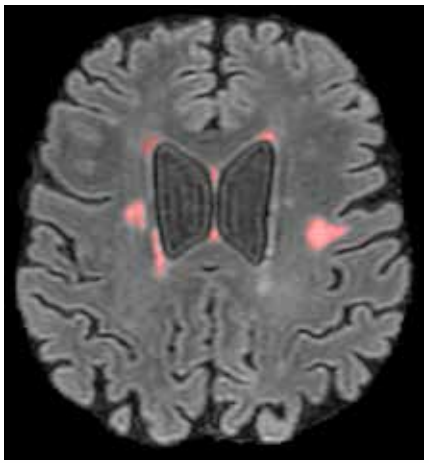
Team 8



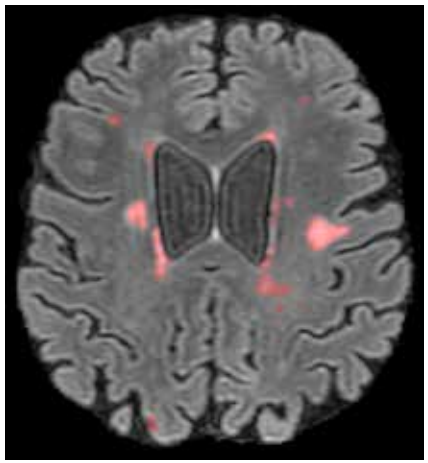
Team 9



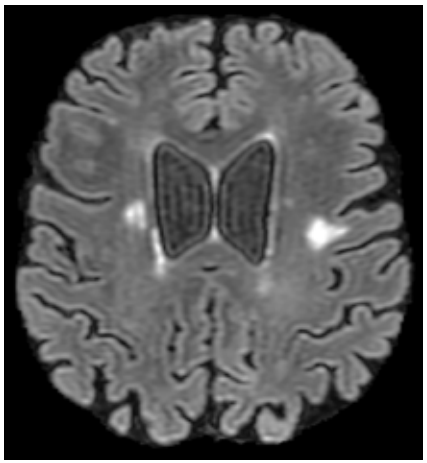
Team 10



Team 11

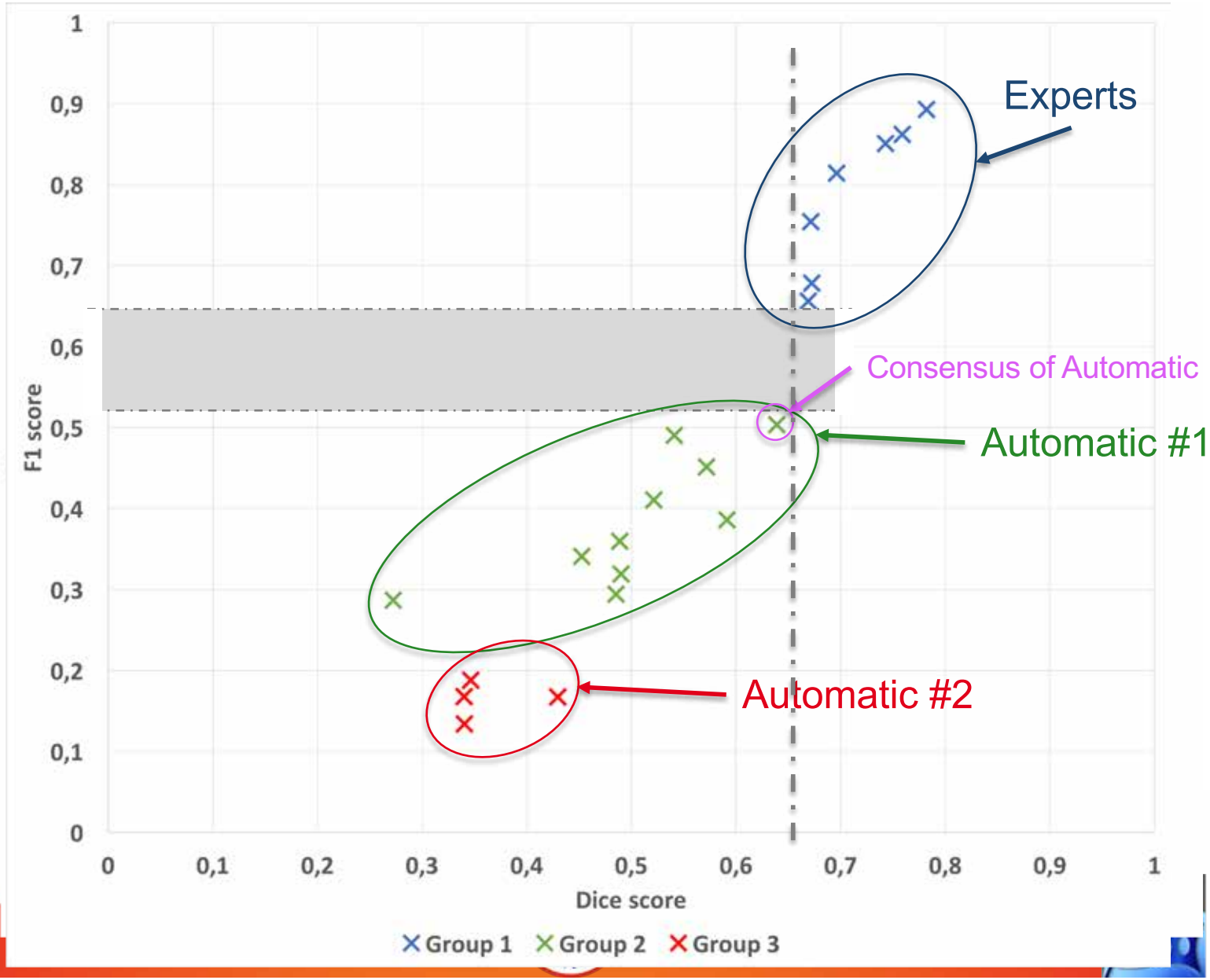


Team 12

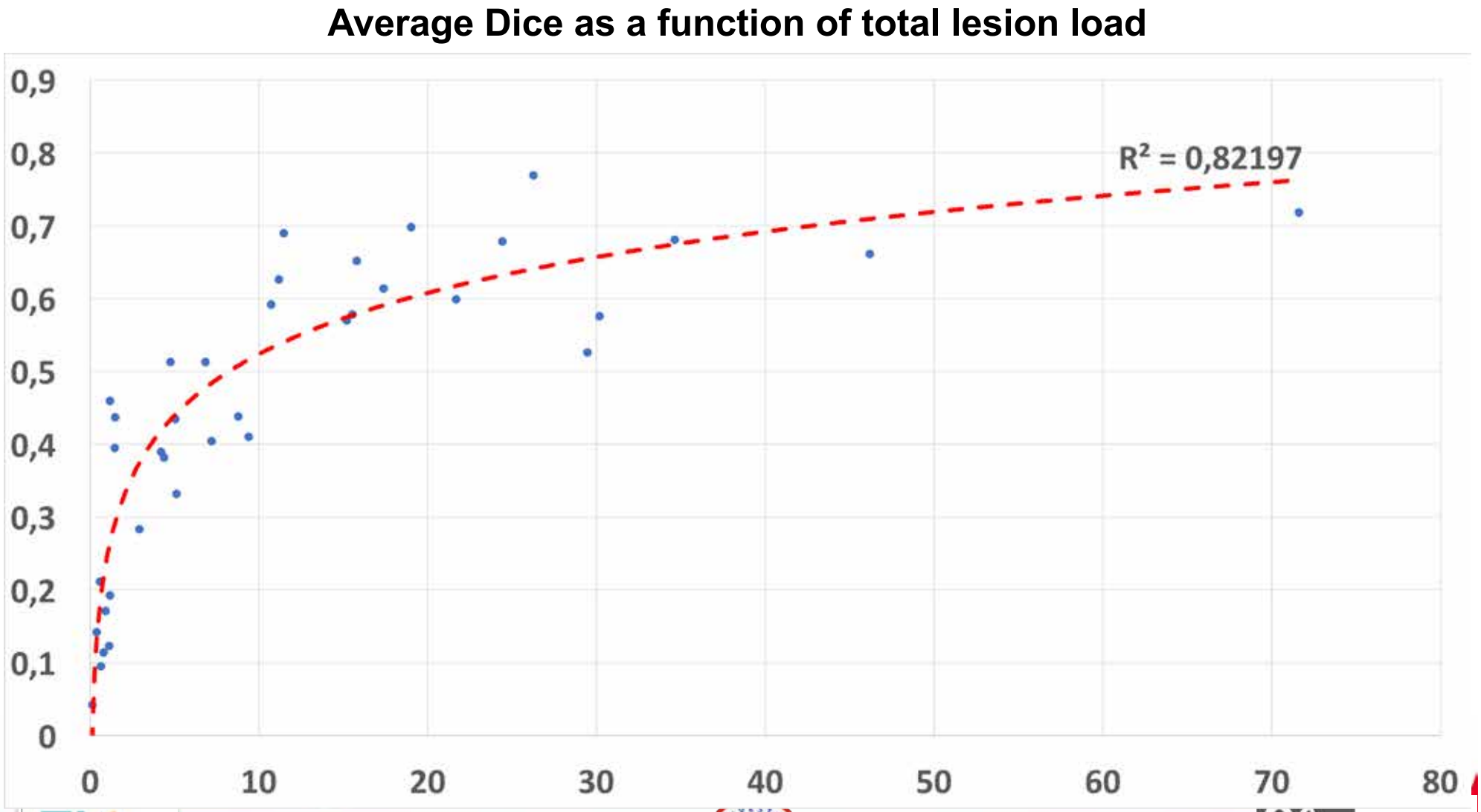


Team 13

Groups of methods : Comparison to Experts



Segmentation performance *vs* lesion load



Take home messages from the challenge

- Standardized acquisitions necessary for MS
  - Yet differences remain
  - Need for large database with many expert delineations (i.e. big issue in medical imaging)
- Automatic computing platform
  - Great tool for
    - challenges organization
    - Open Science
    - Certification of algorithms (e.g. industrial solutions)
  - Fair comparison → no parameter tuning during test
  - No work from challengers after pipeline integration
- Main results
  - Individual algorithms still trailing behind experts
  - Unknown images lead to more failures

France Life Imaging

OFSEP
Observatoire Français
de la Sclérose en Plaques

INVESTIR L'AVENIR