MS-SEG 2016 Segmentation Challenge: Organization and Results

Olivier Commowick and OFSEP / France Life Imaging

January 30, 2018

Outline

- Challenge organization
- Participation and evaluation metrics
- MS lesions segmentation results
  - Outlier case
  - Per center
  - Comparison to experts
  - Relationship to lesion load
- Discussion

An OFSEP and MICCAI challenge

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

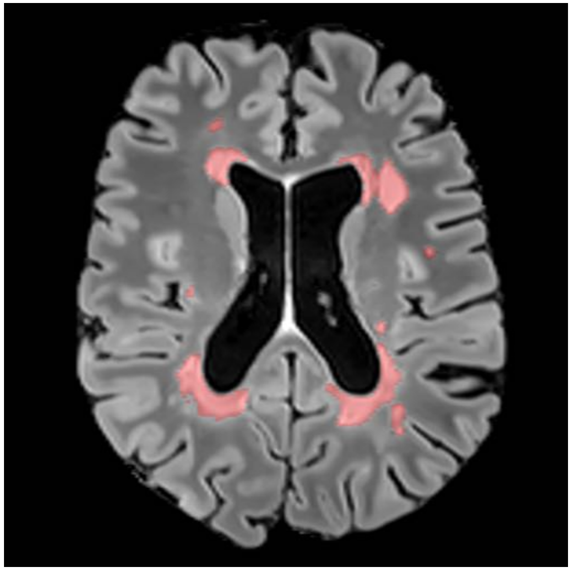
http://www.ofsep.org
Cotton, F., Kremer, S., Hannoun, S., Vukusic, S., Dousset, V., 2015. OFSEP, a nation-wide cohort of people with multiple sclerosis: Consensus minimal MRI protocol. Journal of Neuroradiology 42 (3), 133 – 140.

MICCAI challenge: database

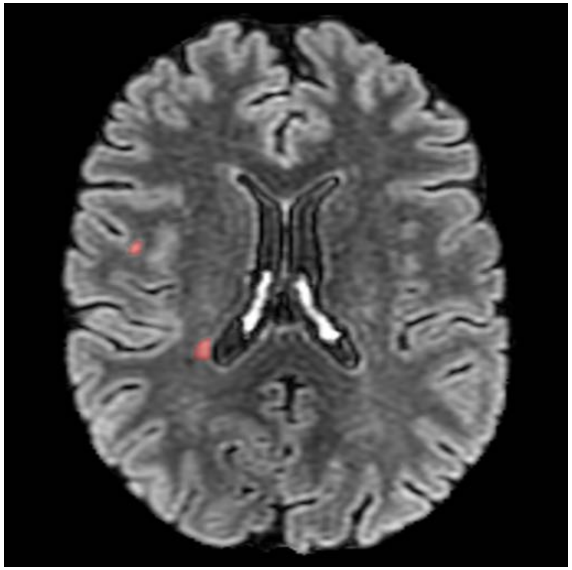
- Challenge data
  - 53 patients from 4 different scanners
  - Modalities: 3DFLAIR, T2/DP, 3DT1, 3DT1-Gado
  - 7 manual segmentations for each patient
- Two datasets drawn
  - Training (open): challengers tune their algorithms
  - Testing (closed): evaluation database

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**

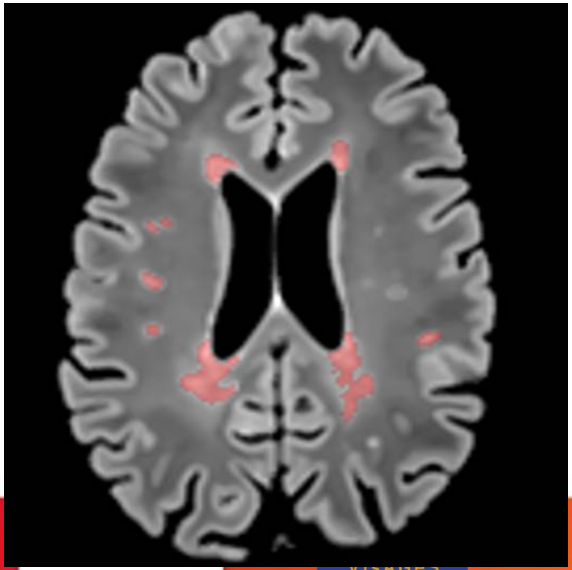
Dataset examples



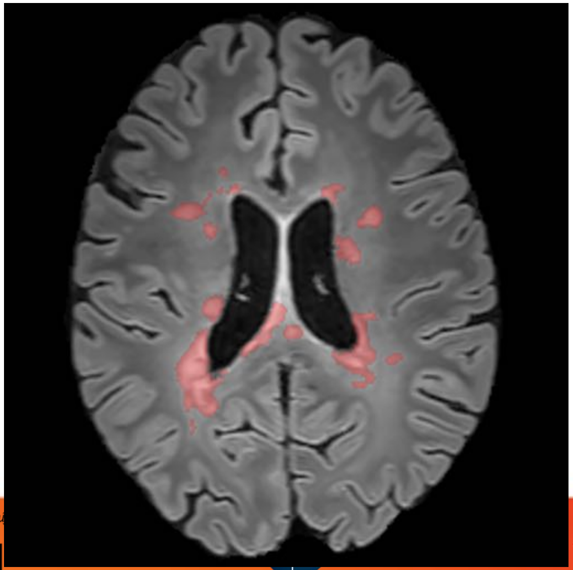
FLAIR from center 01



FLAIR from center 03



FLAIR from center 07



FLAIR from center 08

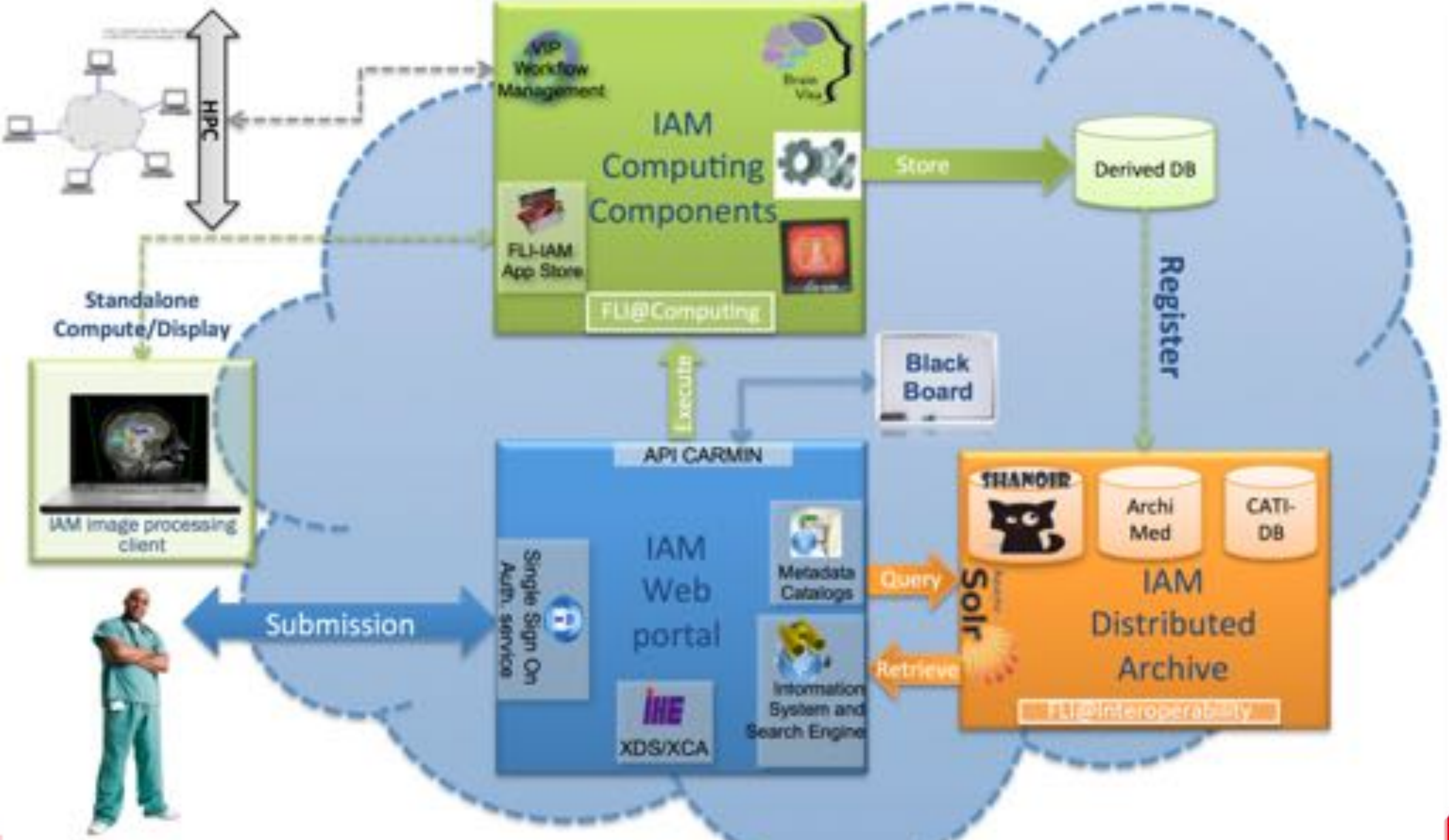
A well defined execution and evaluation framework

- 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) computing platform
  - By FLI 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



Outline

- Challenge organization
- Participation and evaluation metrics
- MS lesions segmentation results
  - Outlier case
  - Per center
  - Comparison to experts
  - Relationship to lesion load
- Discussion

Challenge participations

- Thirteen pipelines including a variety of algorithms
  - Random forests
  - Deep learning
  - Tissue classification approaches

- Training phase: 2 months
- Integration phase: 3 to 4 months
  - Docker packaging and integration help
- Evaluation (independent from challengers): 2 months

Which evaluation? Metric categories

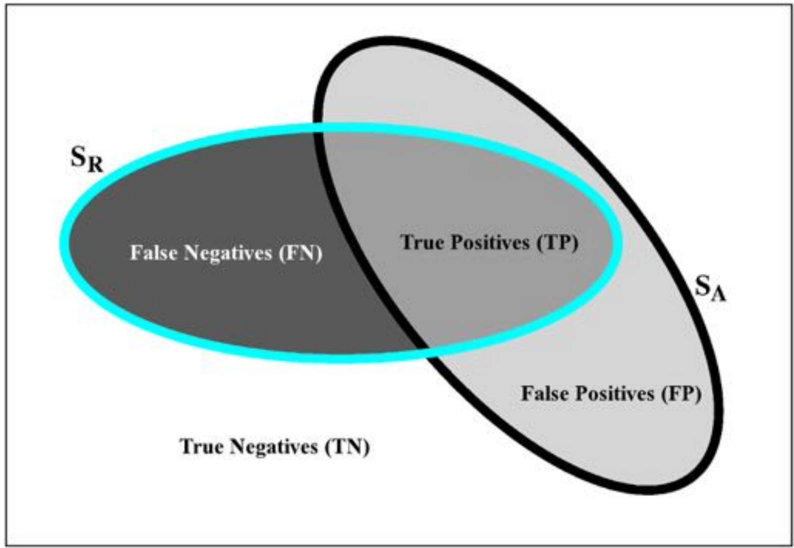
- Evaluation of MS lesions segmentation: tough topic
  - Which ground truth? → LOP STAPLE consensus
  - What is of interest to the clinician?

- Two metric categories:
  - Detection: are the lesions detected, independently of the precision of their contours?
  - Segmentation: are the lesions contours exact?
    - Overlap and surface-based measures

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

Segmentation quality measures

- Overlap measures
  - Sensitivity $D = \frac{TP}{TP + FN}$
  - Positive predictive value $D = \frac{TP}{TP + FP}$
  - Specificity $D = \frac{TN}{TN + FP}$
  - **Dice score** $D = 2\frac{TP}{S_R + S_A}$
- Average surface distance



Detection measures

- Is a lesion detected: 2 criterions
  - Sufficient overlap with consensus
  - Connected components responsible for overlap not too large
- Two quantities measured
  - TP_G: lesions overlapped in ground truth
  - TP_A: lesions overlapped in automatic segmentation
- Metrics
  - Lesion sensitivity and PPV, **F1 score**

Outline

- Challenge organization
- Participation and evaluation metrics
- MS lesions segmentation results
  - Outlier case
  - Per center
  - Comparison to experts
  - Relationship to lesion load
- Discussion

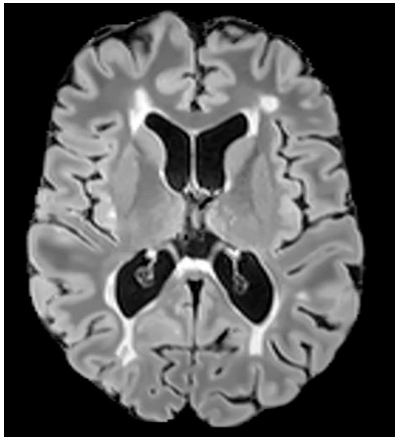
An outlier case study: no lesions

- 5 out of 7 experts delineated no lesion
- Most evaluation metrics are undefined
  - No consensus label
- Two substitution metrics computed
  - Number of lesions detected
    - Number of connected components
  - Total volume of lesions detected
- Both scores are optimal at 0

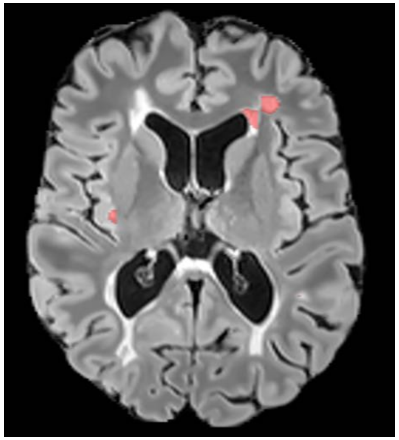
No lesion case results

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

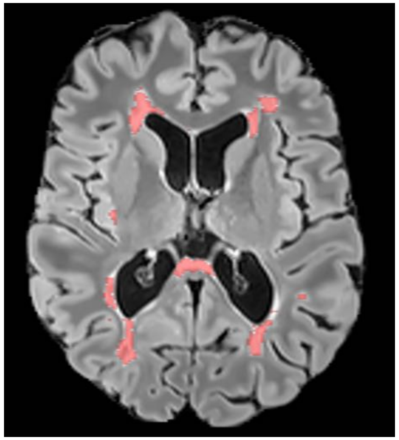
Visual results for center 01



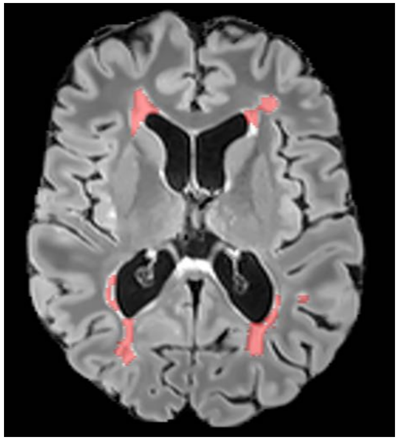
FLAIR



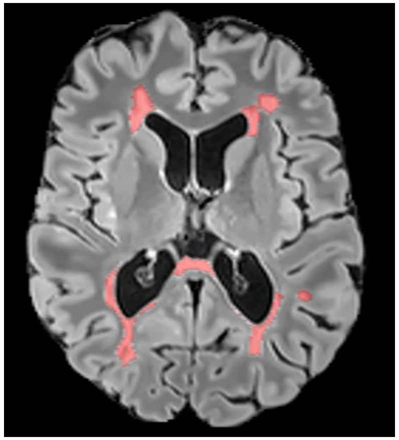
Team 1



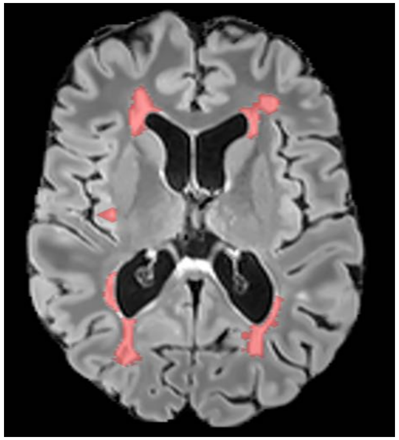
Team 2



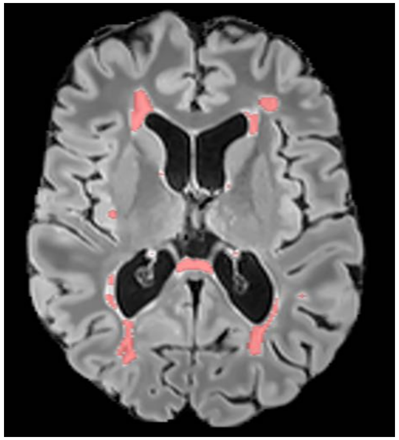
Team 3



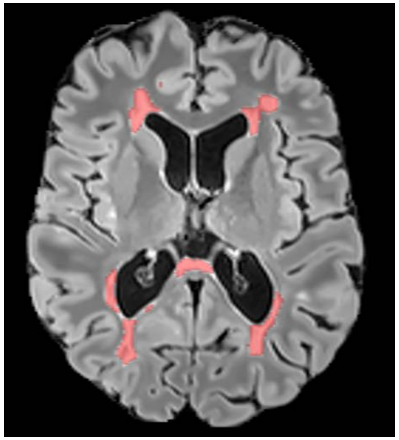
Consensus



Team 4

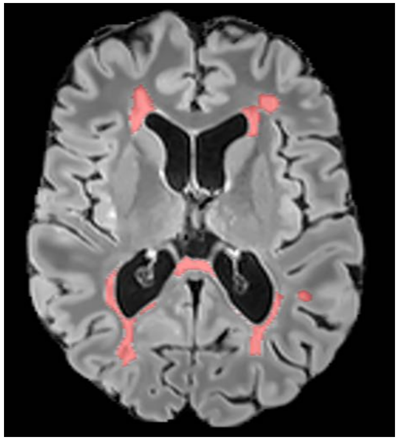


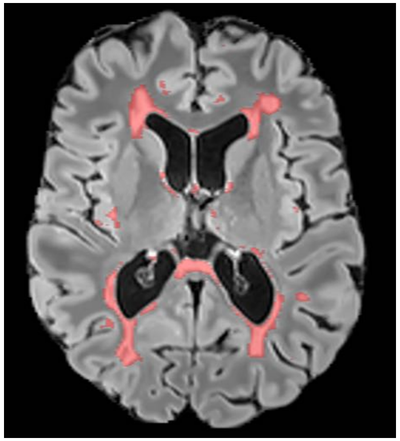
Team 5

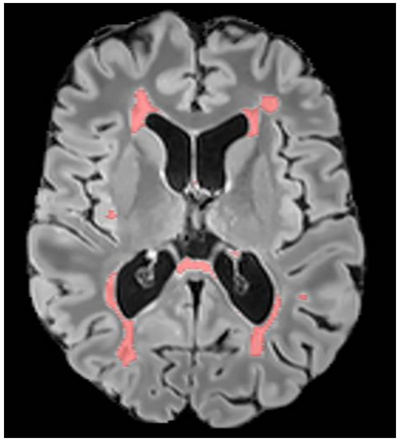


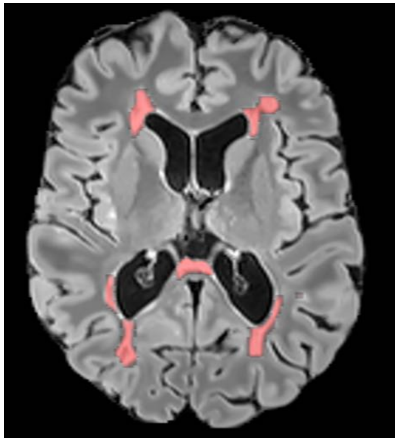
Team 6

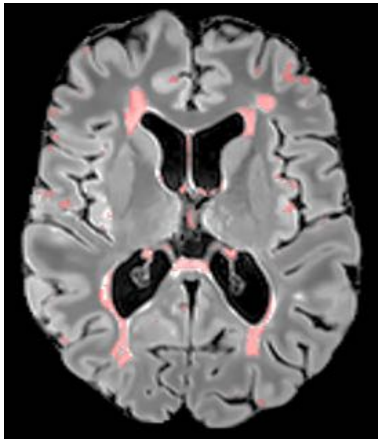
Visual results for center 01

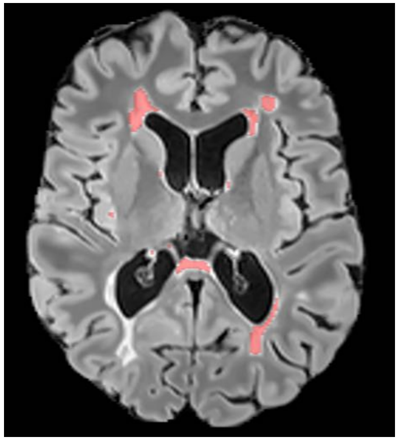

Consensus

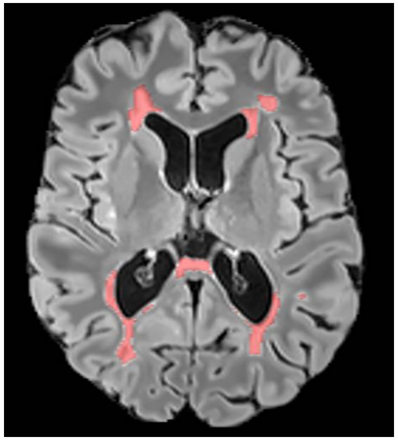

Team 7

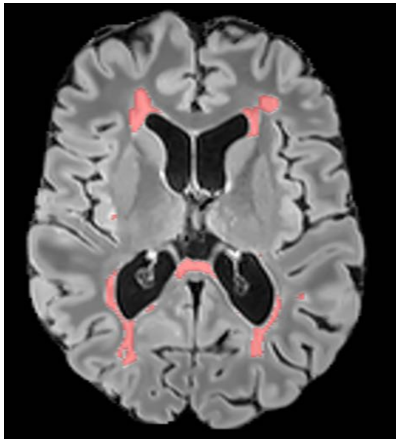

Team 8


Team 9

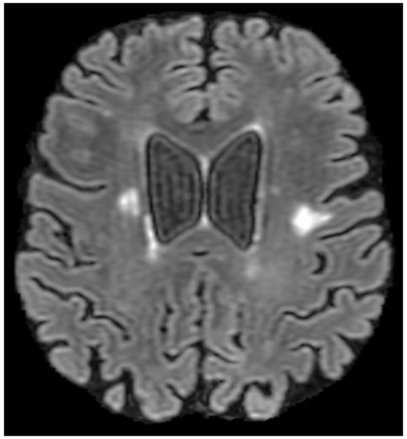

Team 10


Team 11

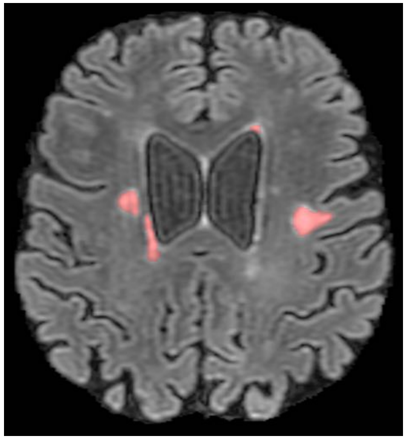

Team 12


Team 13

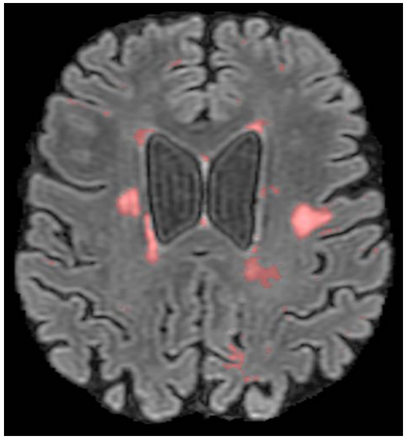
Visual results for center 03



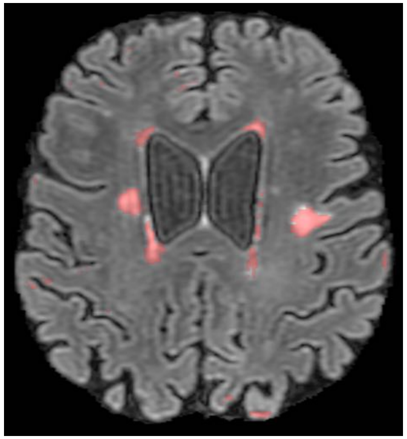
FLAIR



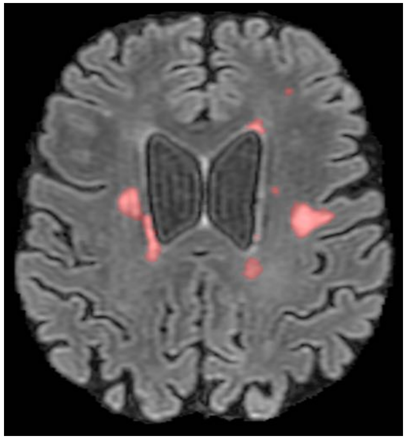
Team 1



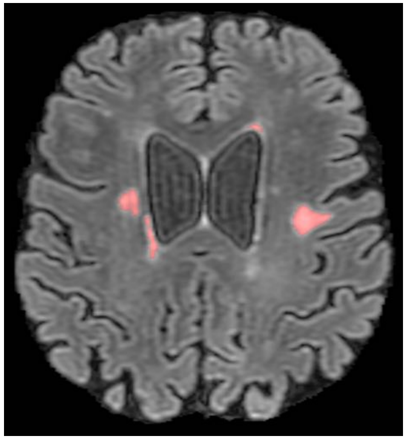
Team 2



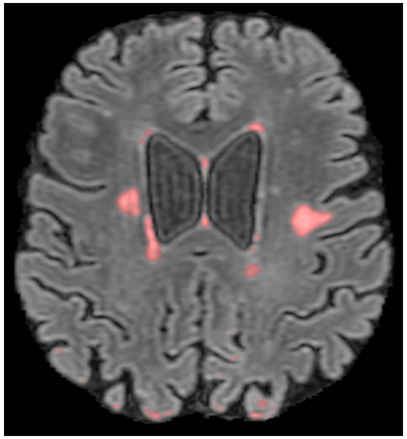
Team 3



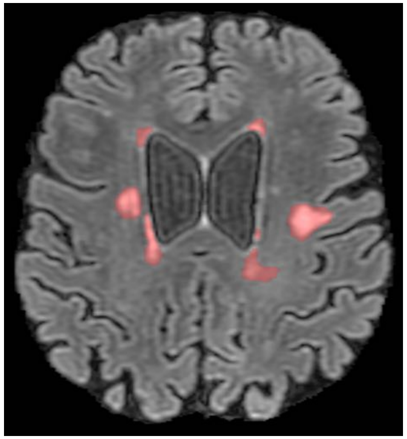
Consensus



Team 4

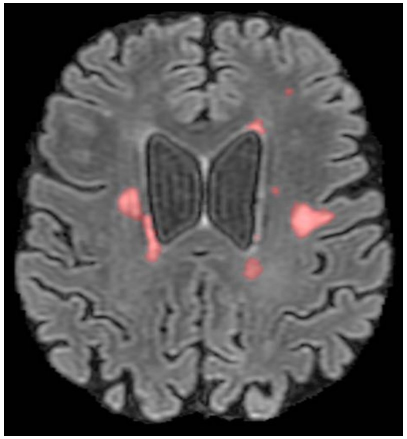


Team 5

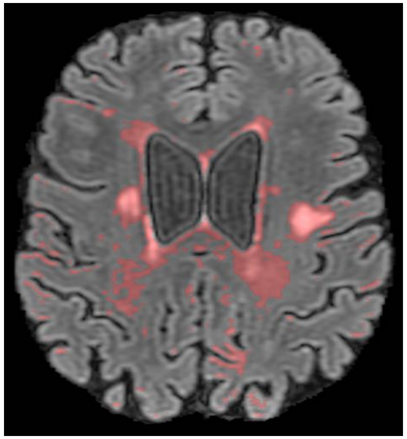


Team 6

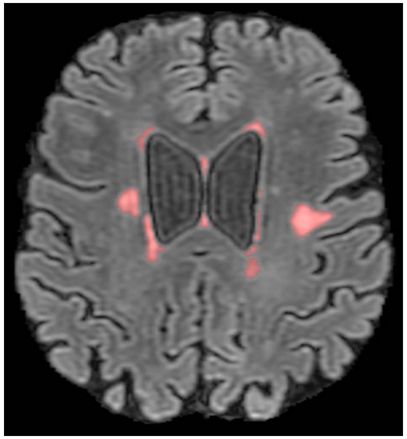
Visual results for center 03



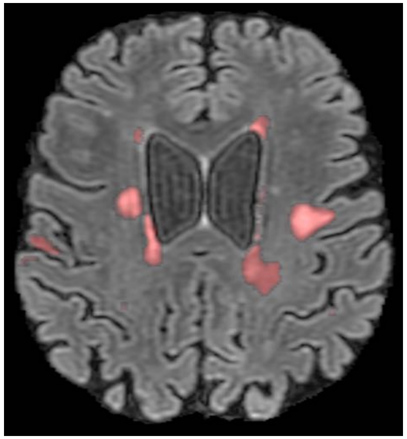
Consensus



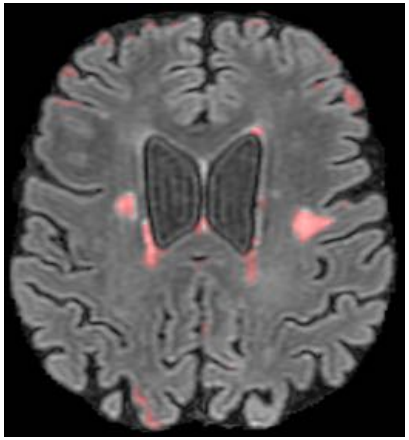
Team 7



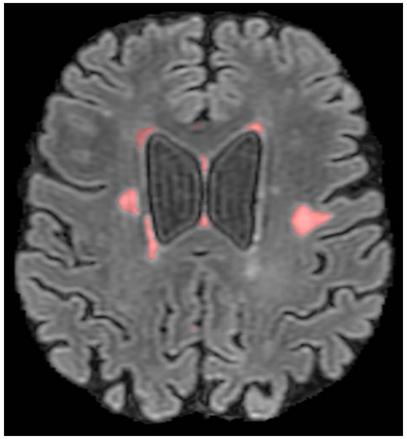
Team 8



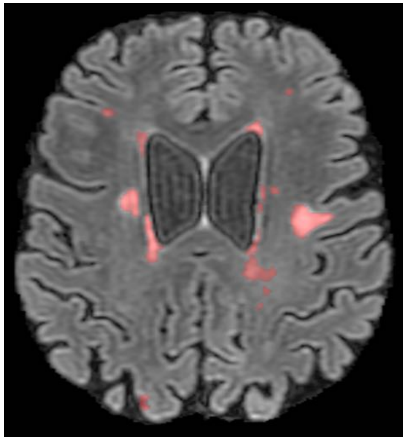
Team 9



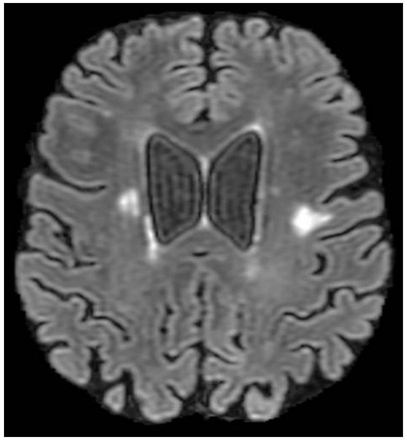
Team 10



Team 11

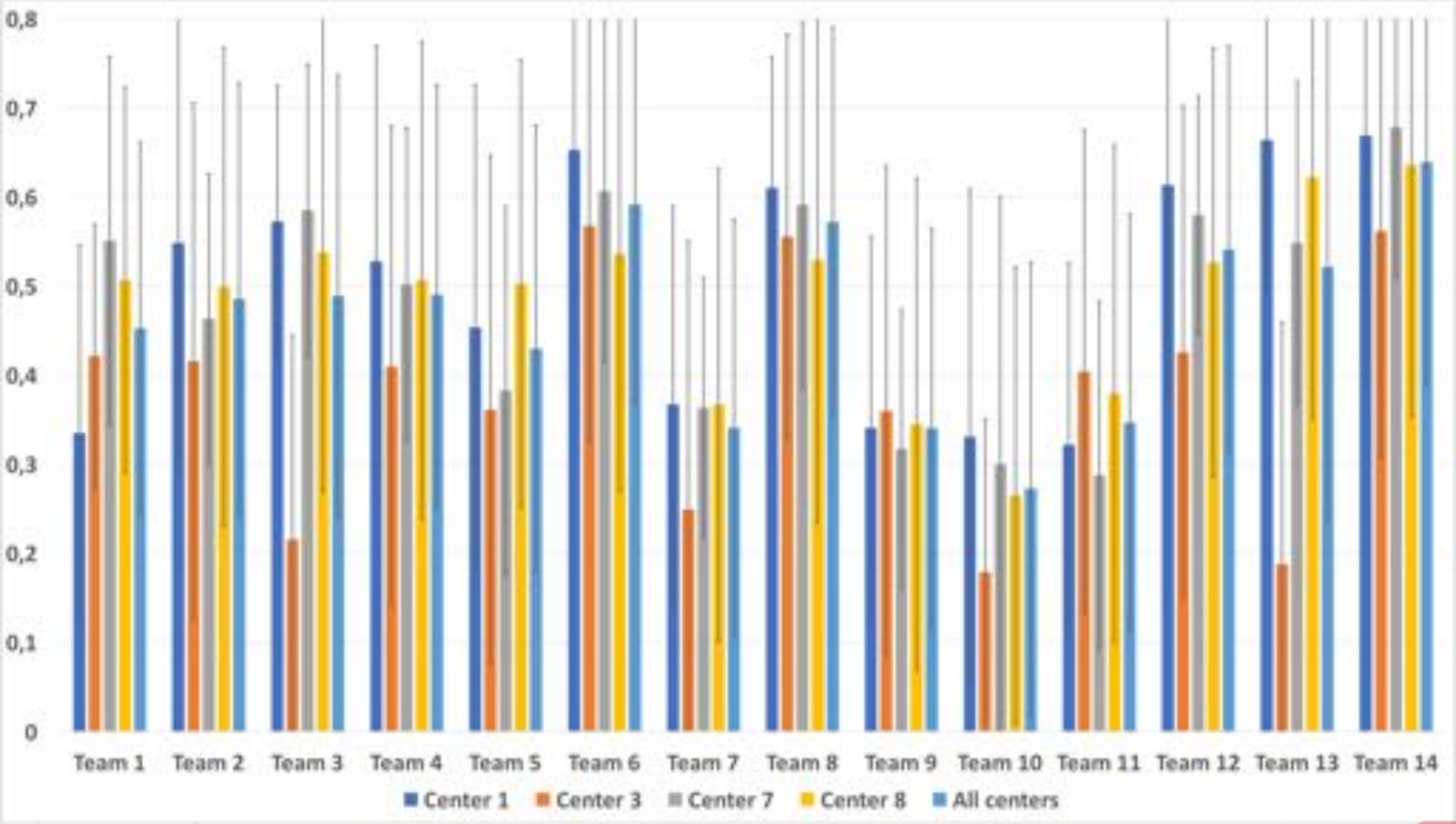


Team 12

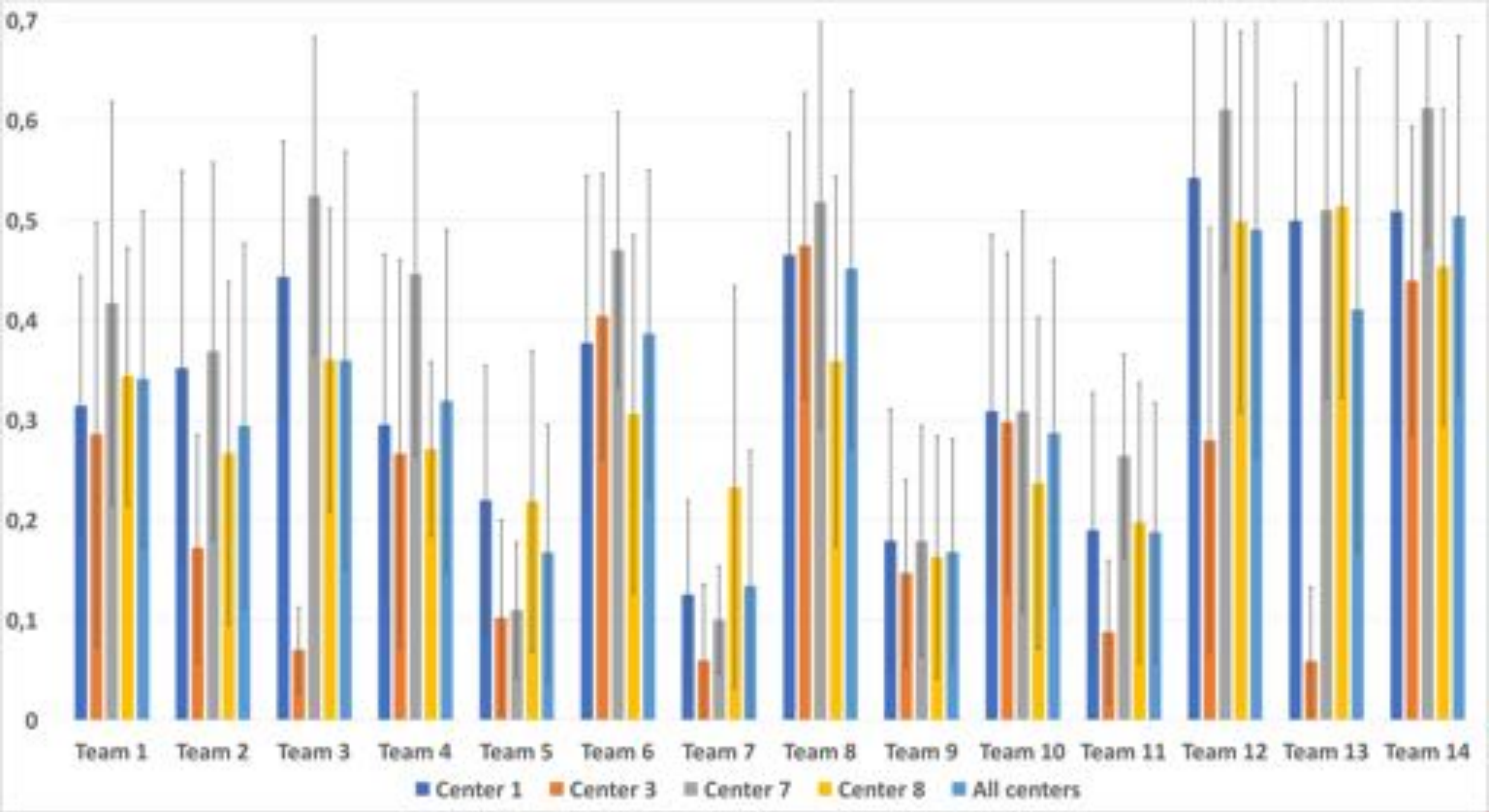


Team 13

Segmentation scores per center



Detection scores per center



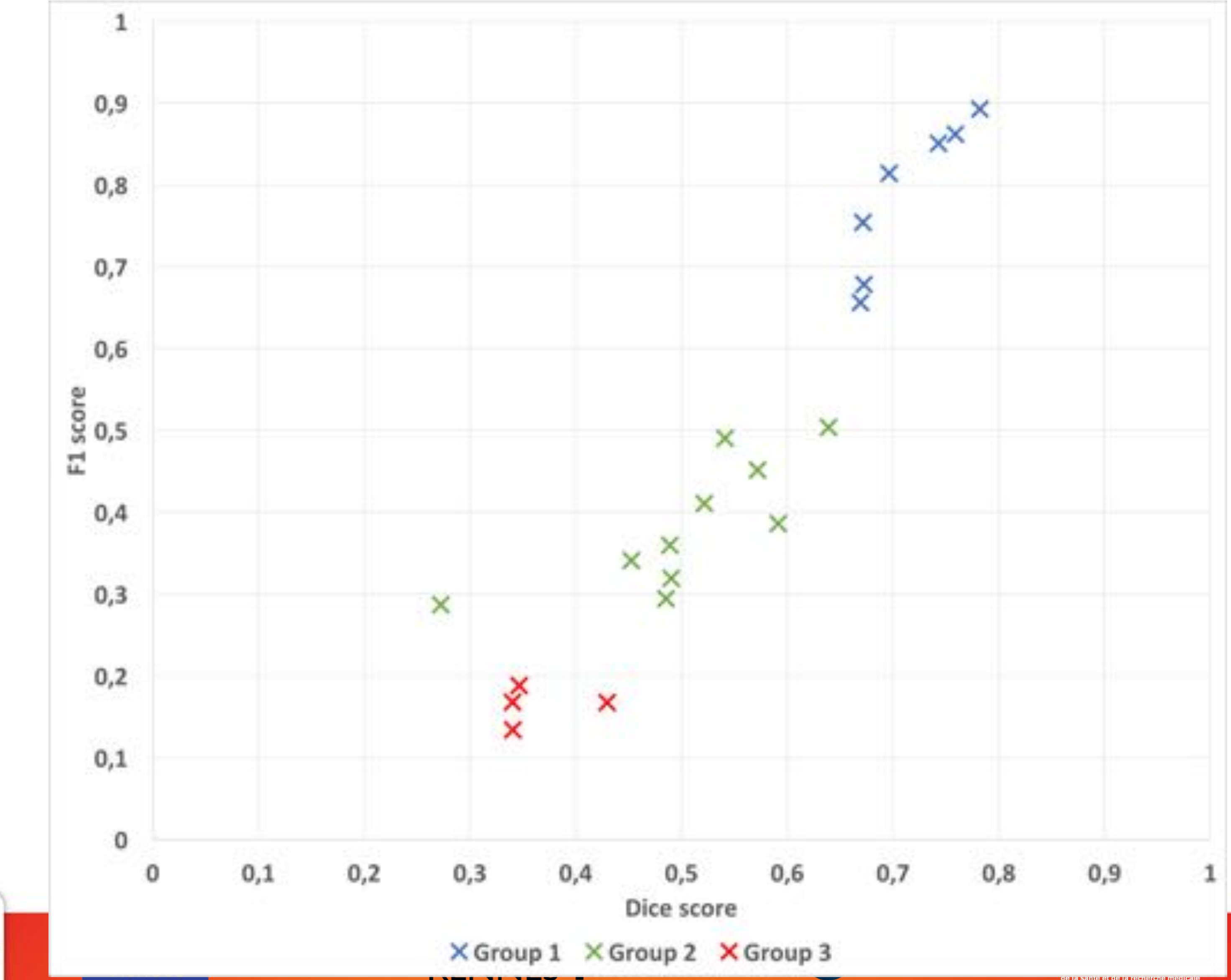
Results comparison to experts

- Are there clusters of algorithms behaving similarly?
  - Clustering from pairs of average measures
    - Surface distance, Dice, F1 score
  - Need to account for variability in measures

- Spectral clustering on experts and methods
  - Calvo & Oller distance to construct affinity matrix
  - Clustering into three groups

Calvo, M., Oller, J., 1991. An explicit solution of information geodesic equations for the multivariate normal model. Statistics and Decisions 9.

Results comparison to experts

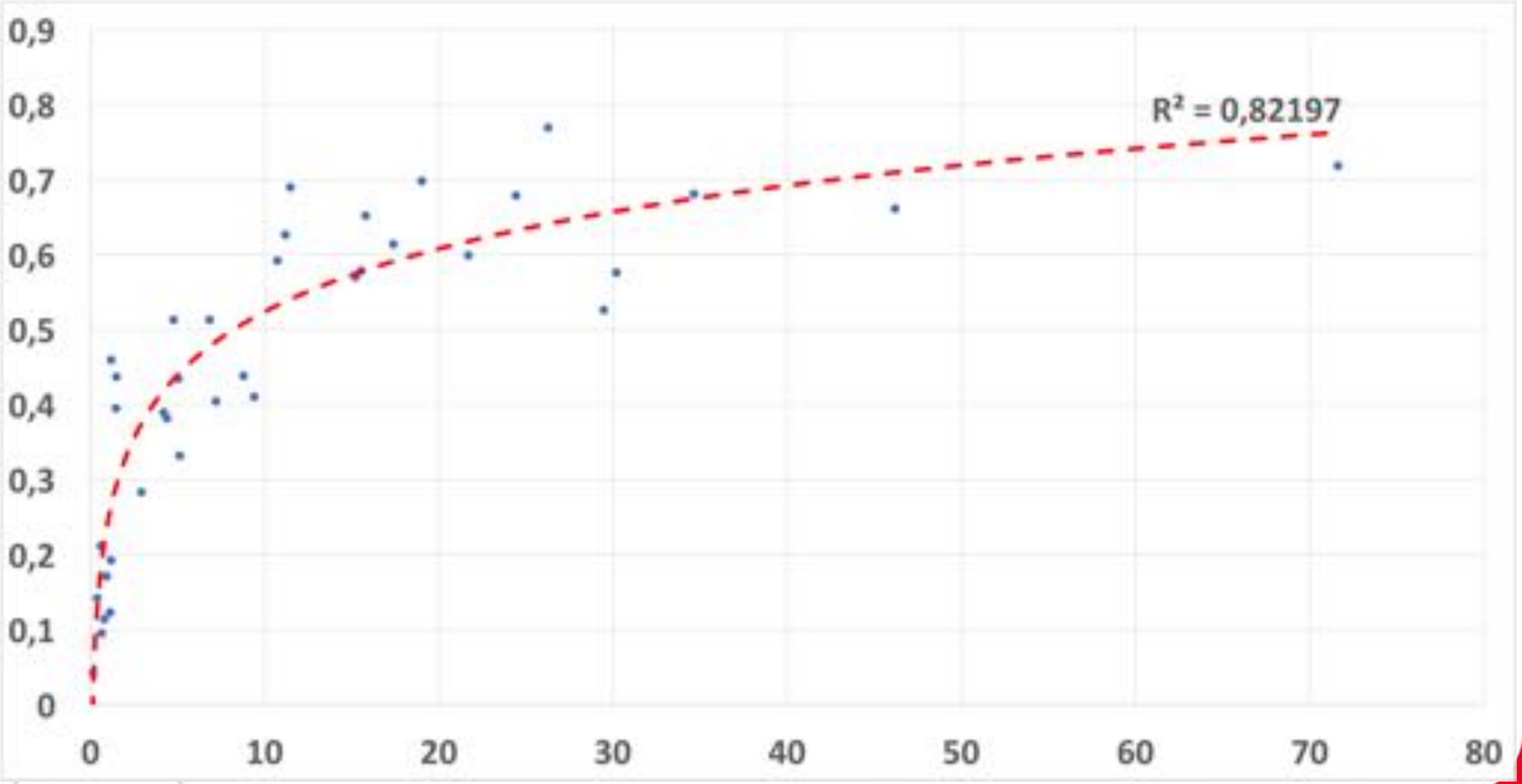


Results comparison to experts

- Segmentation performance
  - “Best” expert: 0.782
  - “Worst” expert: 0.669
  - “Best” pipeline: 0.591
- Detection performance
  - “Best” expert: 0.893
  - “Worst” expert: 0.656
  - “Best” pipeline: 0.490
- All pipelines rank below experts in both categories

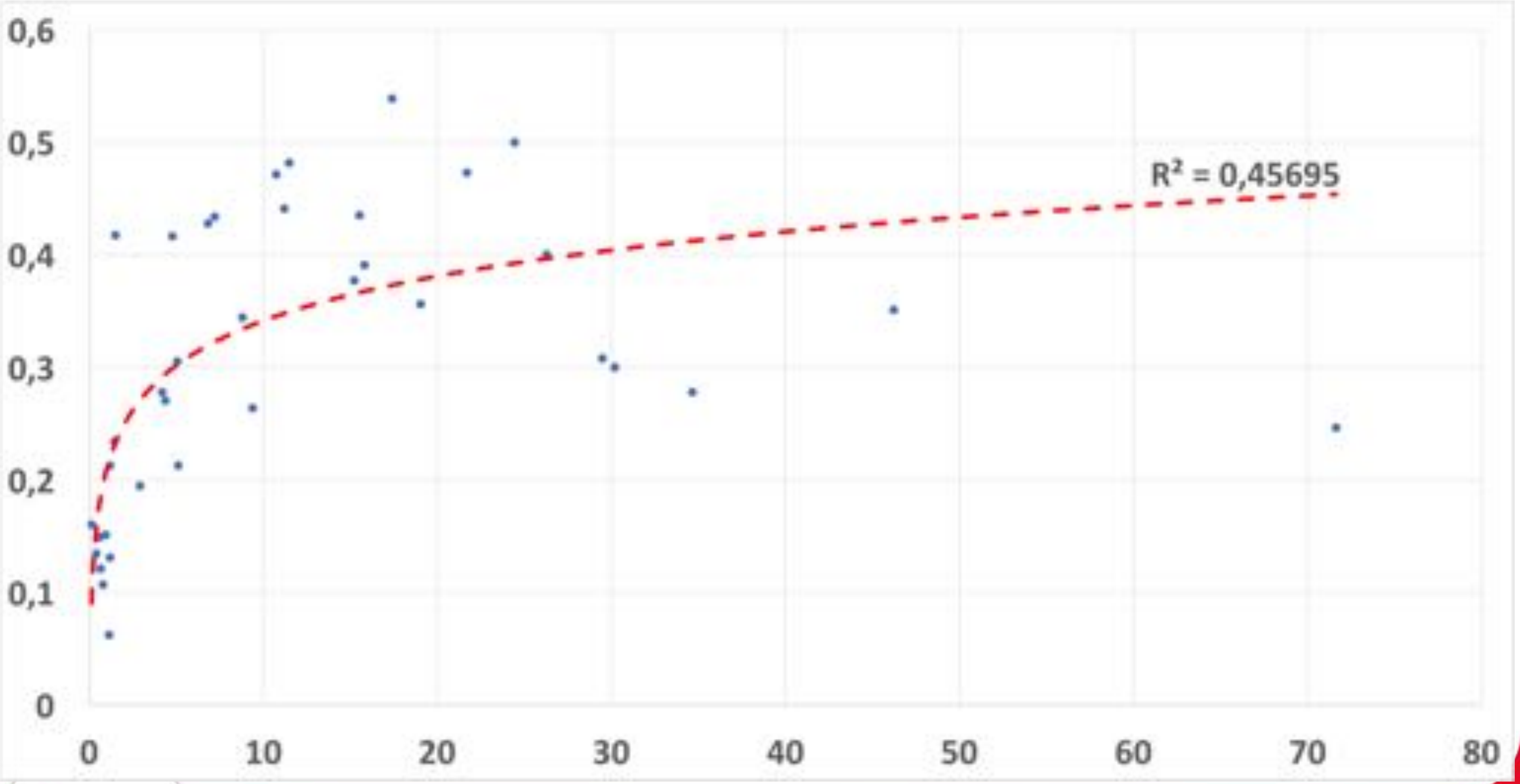
Segmentation performance vs lesion load

Average Dice as a function of total lesion load



Detection performance vs lesion load

Average F1 score as a function of total lesion load



Outline

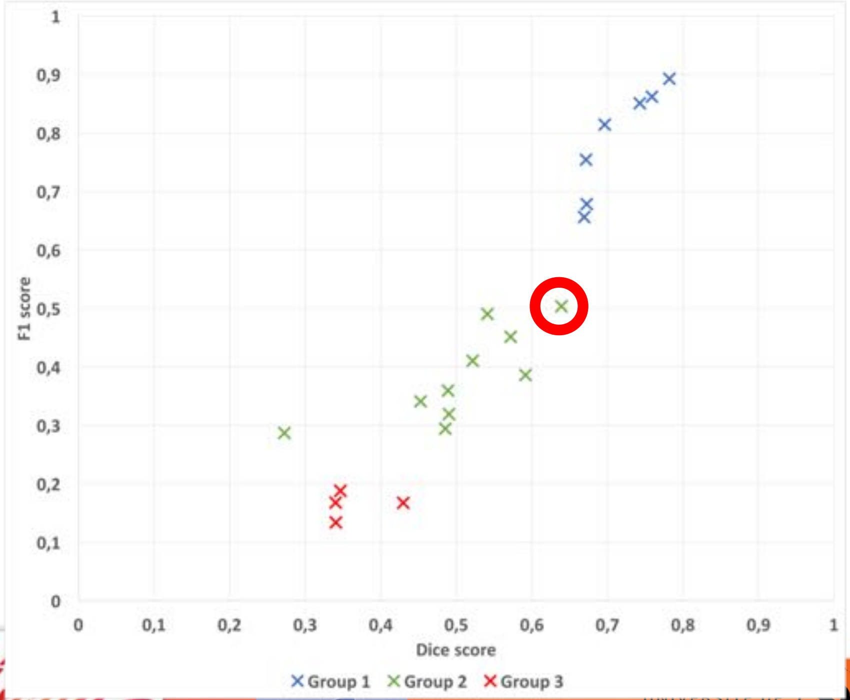
- Challenge organization
- Participation and evaluation metrics
- MS lesions segmentation results
    - Outlier case
    - Per center
    - Comparison to experts
    - Relationship to lesion load
- Discussion

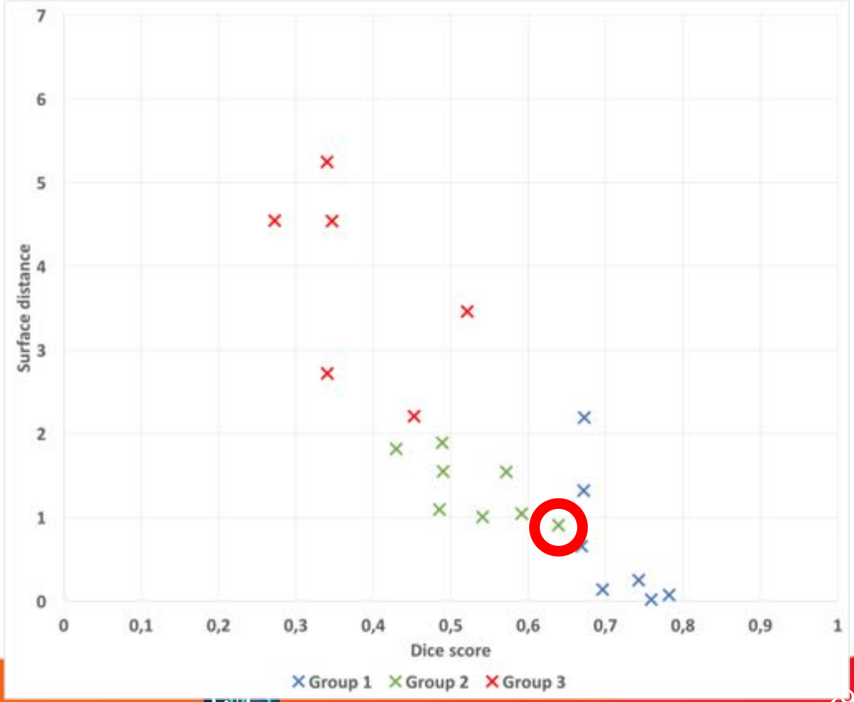
Take home messages from the challenge

- Standardized acquisitions necessary for MS evaluation
  - Yet differences remain
  - Need for large database with many expert delineations
- Automatic computing platform
  - Great tool for challenges organization
  - Fair comparison platform → reduces parameter tuning
  - Platform still opened for evaluation
- Main results
  - Individual algorithms still trailing behind experts
  - Unknown images lead to more failures

Take home messages from the challenge

- Main results (continued)
  - Individual algorithms fail differently
  - Fusion of algorithms improves results


F1 score
Dice score
× Group 1
× Group 2
× Group 3


Surface distance
Dice score
× Group 1
× Group 2
× Group 3