Evaluation criteria for the MSSEG-2 challenge

June 18, 2021

1 Manual segmentations consensus and general evaluation

A total of four manual segmentations are available for each of the 100 MS patients. These segmentations indicate the new lesions, i.e. lesions that appeared on time point 2 and that were not at all present on time point 1. This means that lesions growing or shrinking are not part of the ground truth.

For each patient, we have defined a consensus from the 4 segmentations, using majority voting, after asking an external reviewer to validate disputed lesions (those seen by at most two experts). This consensus, hereafter called ground truth, will be used as the reference against which the automatic segmentations will be evaluated with respect to the experts. The binary consensus maps will be used for the evaluation itself, divided in two categories: segmentation evaluation and lesion detection evaluation.

For each category, a separate ranking of the methods will be performed based on the selected metric. The ranking will be computed as follows:

- Compute the rank of each algorithm for each patient
- Compute the global ranking as the ranking of mean rankings per patient

Finally, two more notes. In addition to those rankings, all metrics will be provided to everyone after the challenge. All metrics mentioned here (apart from those of Section 4) will be computed using the "animaSegPerfAnalyzer" tool in Anima (see the documentation for more details on how to use it), which may be used by the challengers during the training phase. The exact command used will be the following:

```
animaSegPerfAnalyzer -d -l -s -S -i testedSegmentation.nii.gz
-r referenceSegmentation.nii.gz
```

2 Detection evaluation

Evaluation of the detection of new lesions is the most important metric for this challenge. We wish to evaluate in this category how many new lesions have been (in)correctly detected, independently of the precision of their contours. To do so, we will use a metric that was created for the 2016 MICCAI challenge: the detection F1 score. For this task, we first start by computing the connected components of the ground truth G and tested segmentation A and remove all those lesions that are smaller in size than 3 mm^3. We then get label images $\tilde{G}$ and $\tilde{A}$ where each label denotes a specific lesion. The follow-up relies on the redefinition of usual true positive notion to the scale of lesions. For a precise definition of this, see the Scientific reports publication [1]. As for the 2016 MICCAI challenge, we will use the following parameters in that algorithm to evaluate detected and not detected new lesions: $\alpha = 10\%$, $\gamma = 65\%$, $\beta = 70\%$.

From the number of lesions M and N respectively in $\tilde{G}$ and $\tilde{A}$, and TP_G the number of lesions among the M lesions in the ground truth $\tilde{G}$ that are correctly detected by $\tilde{A}$, and TP_A, the number of lesions among the N lesions in the automatic segmentation $\tilde{A}$ that are correctly detected by $\tilde{G}$, the following detection metrics will be computed:

- Lesion sensitivity, i.e. the proportion of detected lesions in $\tilde{G}$: $Se_L = \frac{\text{TP}_G}{M}$
- Lesion positive predictive value, i.e. the proportion of true positive lesions inside $\tilde{A}$: $P_L = \frac{\text{TP}_A}{N}$

The F_1 score is then computed from Se_L and P_L as: $F_L = \frac{2Se_L P_L}{Se_L + P_L}$. This global score will be used for ranking the algorithms. In addition, we will also provide local versions of these metrics (not used for ranking) for sub-parts of the brain (cerebellum, brainstem, lobes of the brain, juxta-cortical lesions). If possible at the time, we will finally try to have the external reviewer review lesions detected by several algorithms but not by the experts to evaluate the capability of the algorithms to provide results not seen by experts.

3 Segmentation evaluation

For each of the experts and automatic methods, we will first quantify how much the new lesions segmentation is precise using an overlap-based metric. This overlap-based metric will be used for scoring and ranking methods. This metric is the Dice score: $D = 2\frac{A \cap G}{A + G}$, where A denotes the evaluated segmentation and G the ground truth. Here notations mean taking the cardinal of each set of voxels.

In addition to this Dice metric, we will also provide other metrics for the sake of a complete evaluation, but these metrics will not be used for ranking. These other metrics include overlap and surface based metrics:

- Positive predictive value: $P = \frac{A \cap G}{A}$
- Sensitivity: $Se = \frac{A \cap G}{G}$
- Specificity: $Sp = \frac{B - A \cup G}{B - G}$, where B denotes the entire image
- Mean surface distance $S = \frac{\sum_{i \in A_S} d(x_i, G_S) + \sum_{j \in G_S} d(x_j, A_S)}{N_A + N_G}$, where d denotes the minimal Euclidean distance between a point of one surface and the other surface, N_A and N_G denote the number of points of each surface.

4 Metrics for images with no new lesions

The training and testing databases include images that do not have any new lesions. This is normal as we wish to have cases as realistic as possible and, in real clinical cases, many patients do not have any new lesion. However, this would lead to undefined metrics if we follow the previous definitions. For these specific cases, we will use, in replacement of the previously mentioned metrics, two simple metrics:

- Number of new lesions detected by the algorithm. This will checked using Anima tool animaConnectedComponents with default parameters.
- Volume of new lesions detected by the algorithm. This will be performed by simpling counting the number of voxels in the segmentation and multiplying by the voxel volume.

For both metrics, the optimal value is 0. The cases with no new lesions will be considered in a specific ranking separated from the other cases. These metrics will automatically be output by animaSegPerfAnalyzer if the ground truth is empty.

References

[1] Olivier Commowick, Audrey Istace, Michaël Kain, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1):13650, 2018.