**Unified Segmentation with Lesion, validation.**

**Big issue:**

How can we validate the results obtained with our proposed USwL approach?

**Data at hand:**

Assume we have access to some imaging data from patients.

For each patient we should have at hand:

* The MR structural images to be segmented/normalized. This can be standard structural T1, T2 and/or PD weighted MRIs or multiple quantitative MRIs (typically R1, R2\*, MT and A).
* A mask of the lesion area(s) in binary format. This one could have been produced by an expert or an automatic procedure.

**The output from USwL:**

After being processed we should thus have the posterior probability map (ppm) to be the ‘lesion’ but also the ppm for the other tissue classes (GM, WM & CSF) and the intra-cranial volume (ICV) map.

**Validation approach(es):**

With the input and output images, how can we assess the USwL procedure? The goal is not unique neither simple; we want to show that we get “better” output segmentation (lesion ppm, i.e. ppm-LES) than what is input (original lesion mask, i.e. msk-LES).

Here are a few ideas, in random order:

1. Calculate some “overlap index”, a.k.a. “inter-rater agreement”, between the ppm-LES and msk-LES. We would limit ourselves to voxels within the intra-cranial volume. This corresponds to comparing 2 vectors, one with values within [0 1] and another one with 0/1. Here are a few possible measures:
   * A simple correlation between the 2 vectors or the percentage of agreement. Pro’s: simple; con’s: not well behaved because of the unbalancedness (many more healthy than lesioned voxels).
   * Kohen’s Kappa (<https://en.wikipedia.org/wiki/Cohen's_kappa>) is more robust as it accounts for the random agreement.
   * Dice coefficient (<https://en.wikipedia.org/wiki/Dice%27s_coefficient>)

QS = \frac{2C}{A + B} = \frac{2 |A \cap B|}{|A| + |B|}

or Jaccard index (<https://en.wikipedia.org/wiki/Jaccard_index>)

J(A,B) = {{|A \cap B|}\over{|A \cup B|}} = {{|A \cap B|}\over{|A| + |B| - |A \cap B|}}.

are standard measures of similarity between finite sample sets with values in [0 1]. Both are related by =D/(2-D) and =2J/(1+J) where D is Dice coefficient QS.

These measures could lead up to 5 numbers.

Further question: at which level to threshold the ppm-LES? Should we agree on this in advance and fix it, say at .5? Or what about looking at how the measure values change over a range of thresholds?

1. Look at the intensities in the structural images of the voxels on the border of the msk-LES. USwL should improve lesion segmentation especially around the contour, as manual or semi-automatic segmentation is more likely to be 1-2 voxels too large/short while the output USwL relies on voxel intensity and should improve the contour. We can compare the mean and std values of the intensities of the structural MRIs right inside/outside the msk-LES, accounting or not for the ppm-LES. I would expect that the intensities are more homogenous when relying on ppm-LES.

Similarly to the previous point: at which level to threshold the ppm-LES? Should we agree on this in advance and fix it, say at .5? Or what about looking at how the values change over a range of thresholds?

1. With data from a patient, add/remove some lesion from the original msk-LES:

* add an erroneous (small) blob in the msk-LES and see how the ppm-LES looks like. This will assess the *specificity* of the USwL procedure, i.e. “can USwL be mislead by wrong priors?”
* remove a true (small) blob in the msk-LES and see how the ppm-LES looks like. This will assess the *sensitivity* of the USwL procedure, i.e. “can USwL detect lesions that are missing from the priors?”

Another possibility is to create a data set with lesion (so we know the gold standard) and to proceed as here above to check the specificity and sensitivity of USwL.

1. Check if the normalization of the whole brain is better with USwL than simply masking the lesions. How to assess this? Here are a few ideas:

* Look at the GM overlap across the subjects. To do this one could rely on a paired t-test of the ppm-GM, with USwL or “simple masking”?
* Check the voxel intensities of voxels in tissue class(es).

Other ideas?

**Comments:**

#1 will show that we recover about the same thing as we put in, so that’s not really useful to us. And by varying the threshold we could see how the ppm-LES threshold affects these overlap measure but we still don’t know what’s right/wrong as the msk-LES is not 100% correct.

#2 would really indicate an improvement in contour segmentation. And if we vary the ppm-LES threshold we can possibly find the optimal threshold value. One possible issue: one needs to use sufficiently large blobs in order to have voxels inside/outside the blob.

#3 assess the robustness of the approach when there is erroneous presence/absence of lesion mask in msk-LES.

#4 is straightforward but unlikely to work since the effect of “simple masking” vs USwL is bound to happen all over the place in the brain. One would rather look for an increase/difference in variance of the GM between the USwL and “simple masking”. Though this may not be sufficient…

**Conclusion:**

Overall #1 is kind of necessary and just useful to reassure ourselves that USwL does not do crazy things.

#2 and #3 are good criteria. #2 gives values based on image intensities and is not subjective. #3 tells us how much we can be wrong to begin with and how much we can trust the output results.