Sunday, July 13, Copenhagen
HarP Certification meeting

Attendees:
Giovanni B Frisoni, Clifford Jack, Marina Boccardi, Simon Duchesne, Martina Bocchetta, Heather Snyder, James Hendrix, Meredith McNeil, Nick Fox, Josephine Barnes, Marc Modat, Ged Ridgway, Christos Davatzikos, Marco Lorenzi, Joyce Suhy, Chahin Pachai, Andy Simmons, Louis Collins, Catherine Gray, Nicola Amoroso, Zuzana Nedelska, Michela Pievani, Maura Parapini, Paola Palazzo, Enrica Cavedo,

Remotely:
Chen Gennan, Ron Piersen

Aim of the meeting:
To present the project for an Alzheimer's Association (ALZ) platform allowing certification of the harmonized protocol (HarP) compliance for automated hippocampal segmentation algorithms.

Discussion:
Giovanni Frisoni (GBF) describes the background of the Harmonized Protocol (HarP) for Manual Hippocampal Segmentation project, from its rationale: dysfunctional heterogeneity of volume estimates by different protocols, up to 2.5folds for a normal hippo, through the different phases: operationalization (the Lego block exercise) of variability among most used protocols, Delphi panel selection of landmarks, drafting of the user manual describing landmarks and segmentation procedure, production of benchmark image and infrastructure (Certification Platform) for manual segmentation certification, and validation results.
GBF underlines the very high absolute ICC values, and the very low coefficient of variation due to tracer, and how these excellent results support the use of the HarP as a standard procedure for hippocampal volumetry for clinical trials and for clinical diagnosis.

GBF then presents the so-called "Expansion" project, aimed at producing additional benchmark labels from 135 different ADNI subjects balanced by main confounders. These labels will serve the training of automated hippo segmentation algorithms. So far, a preliminary set of 100 labels has been released for public use, but the final set will be released shortly.

GBF then reports a set of published algorithms for automated hippocampal segmentation. These algorithms resemble the same segmentation heterogeneity of the different protocols and their employment for standard use requires the same compliance to a single standard as obtained for manual segmentation. Such compliance will need to be certified in order to allow standard use of such algorithms. Compliance and certification need to be consistent with current Europe and US health care and clinical trials requirements (slide 16).

The ALZ HarP Certification will aim to certify that a specific release of a specific algorithm can segment the hippocampus consistently with the HarP under specific circumstances (e.g. image quality). It will NOT certify that an algorithm is compliant with any regulatory directive, nor will it replace any EMA or FDA directive for lawful marketing of medical device. The ALZ Certification Platform will be chaired by the ALZchairperson, managed by the co-PIs GBF and Cliff Jack, and by a set of experts (Nick Fox as clinical neuroscientist, Paul Thompson as computer scientist, and physics/engineers). Additional personnel will be selected as program

develops. Personnel with significant conflict of interest with private companies will not be selected.

The infrastructure is based on the Certification platform already set for manual segmentation. The platform needs to be implemented with the "expansion" labels and with tests and thresholds for the comparisons of algorithms versus the standard. Metrics will be agreed upon by experts of automated segmentation, and thresholds defined by the ALZ Expert committee based on the manual segmentation performance. Laval University (Simon Duchesne) is willing to share the platform to the ALZ, due to potential perceived conflict of interest in the company he runs with Louis Collins. Cliff Jack is comfortable with Simon running the platform with clear transparency of his conflict to potential users, due to his know-how. GBF agrees with this, but says that if anyone among potential users is not comfortable with this, a different institute will be charged with this task.

The platform will allow users to submit their algorithms, have metrics computed versus test images, and receive a feedback about compliance (certification/advice for improvement/rejection). From the audience, the "advice for improvement" feedback may not be given, and only a pass-or-fail feedback may be available.
On the whole, the ALZ certification platform should be an interface between academia, algorithm developers, regulatory agencies and related working-groups.

Discussion:
Chen: what will happen with different releases of a same algorithm?

CJ answer: different releases will need to undergo *de novo* certification

Simon Duchesne: actually, we do not certify the algorithm. Rather, we certify *its output labels as compliant to the specific testing set.* It is the output labels that are submitted and certified, not the algorithms themselves.

(Christos?): says that the list of algorithms that we reported is limited. It might also include many algorithms that allow for parcellization, and may include hippo segmentation within general parcellization.

Chahin: our plan should specify which is the set of images/labels that will be used to test the algorithms. This should not be released to users. Suggests not to release the last (late) 35 labels and use them for the test phase.

Discussion ensued, suggesting that we may use also the 20 labels of the platform for manual segmentation that we use to test manual tracers. These have never been released publicly. Additional labels from 16 subjects from the Validation phase may be used, choosing among segmentation performed by the "best tracers".

Christos: the test set should be balanced by confounders.

Answer: both the 20-labels and the 16-subjects labels sets are balanced by confounders, but we must check about the 35-subjects set.

Christos: it would also be useful to have different sets for different test sections.

Answer: Nick says that additional images may be required to be segmented, so that the test images are hidden, and users do not know on which images they will be tested.

- Also, the frequency for test access should be defined, otherwise one may repeat the test until certified

- the feedback should only be pass-or-fail, without the "suggestion for improvement". Maybe we may say on how many images the algorithm failed.

- what is the optimal size for testing reliability?
  Ron Pierson uses 30 cases for manual reliability

Ged and Jo Barnes: challenging images may be proposed, even only for the training phase.

Andy: opportunity to introduce negative controls. How can we evaluate segmentations performed on difficult images (movement, bad contrast..)
We can account for performance as carried out on MR following the ADNI standard, so this is out of our scope even if may be relevant in clinical settings.

Provided feedback: rather than only pass-or-fail, some measures (e.g. global Dice) may be provided to be transparent.

1) Think how to integrate null results into final statistic
2) Think of statistic
   Metric threshold per hippocampus
   Metric threshold overall
3) Think of access - block multiple certifications
4) Think of including false negatives
   Images with artefacts
   Images with atypical presentations

Next Steps: (*post-meeting with GB Frisoni, CR Jack, S Duchesne, J Hendrix, H Snyder, M Boccardi*)
The Alzheimer's Association Certification Platform Committee
- Name a chairperson from the Alzheimer's Association.
- Confirm Nick Fox and Paul Thompson as expert members.
- Define the total number of expert members needed for the committee and their roles.
- Nominate additional expert committee members.

Define the infrastructure needs for testing and certification.
- Define a detailed budget for the project to be evaluated by the Alzheimer's Association and the ALZ Committee.
- Define specific tasks to specific groups.

Process for certification:
- Define the rules and process for certification to ensure the integrity of the test, including the training and test images.

Set a date and time for the next team meeting.

Additional key issues (from subsequent discussion – *GB Frisoni, CR Jack, S Duchesne, M Boccardi*)

Increase test set. Unpublished benchmark labels may be useful for the test phase:

Current certification test set (10 subjects) BALANCED FOR CONFOUNDERS
Expansion release (35 subjects) *CHECK BALANCE OF CONFOUNDERS*
Validation test set (16 subjects)(2 timepointes?) BALANCED

However the images corresponding to the above labels can be accessed by those who have a good knowledge of the project. This results in potentially unfair testing conditions with respect to users who were not able to access these images. To guarantee fair testing conditions to all, it is necessary to produce new benchmark segmentations (labels and image codes *will not be published*).

6) Think of charing a fee
Sliding scale - cheaper for small businesses, higher for large businesses
Donations to AA

7) Think of **involving someone from FDA, EMA, Health Canada**….