Variability Issues in Automated Hippocampal Segmentation: A study on out-of-the-box software and multi-rater ground truth

M. Liedlgruber*, K. Butz†, Y. Höller†, G. Kuchukhidze†, A. Taylor†, O. Tomasi‡, E. Trinka†, A. Uhl*
†Department of Neurology, Christian Doppler Medical Centre and Centre for Cognitive Neuroscience,
Paracelsus Medical University, Salzburg, Austria
*Department of Computer Sciences, University of Salzburg, Austria
‡Department of Neurosurgery, Paracelsus Medical University, Salzburg, Austria

***Abstract*—In automated hippocampus segmentation, issues related to ground truth rater variability, subject variability and variability of software segmentation accuracy are investigated in the context of 3 publicly available, out-of-the-box software packages. Ground truth variability among three manual raters is controlled using a majority voting based label fusion scheme and observed subject variability underpins the importance of availability of large scale ground truth.**

***Keywords*-automated hippocampus segmentation, ground truth variability, subject variability**

I. Introduction

The hippocampus is reduced in size in individuals with obesity, diabetes mellitus, hypertension, hypoxic brain injury, obstructive sleep apnoea, bipolar disorder, clinical depression, and head trauma, it is atrophic in mild cognitive impairment and dementia [1], and it is sclerotic in specific subtypes of epilepsy [2]. Since the hippocampal formation plays a core role in memory formation and consolidation [3], pathological reduction in size and other structural pathologies correlates with cognitive decline [4], [1]. The most established application of hippocampus volumetry is the prediction of conversion from normal aging to mild cognitive impairment, and further to Alzheimer disease [5], [6], [7], [8], [9].

Thus, a large variety of techniques and algorithms for automated hippocampus segmentation have been published over the last years, some of them targeted to specific disease or deformation classes (see e.g. [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]). In order to assess these software solutions, the classical evaluation approach requires manually established expert ground truth segmentations, a tedious and time-consuming effort, especially when larger datasets are used. For the state-of-the-art algorithms for automated hippocampus segmentation [20], [21] based on multi-atlas segmentation (MAS [22]) it is not only assessment that requires reliable ground truth, but each involved atlas usually relies on a manually established expert ground truth segmentation.

Therefore, in literature, alternatives have been pointed out recently: [23] introduces a generic learning-based approach based on a set of segmentation features which are trained to predict overlap error and Dice coefficient of a segmentation of unseen data without available ground truth. Crowdsourcing has been identified as a further approach to at least approximate expert label quality by combining several non-expert segmentations [24] and other medical image data label generation tasks [25], [26].

As a consequence of the significant effort in generating expert ground truth segmentations, many studies involve only a single rater groundtruth per MRI volume for hippocampus segmentation accuracy assessment (e.g. [13], [14], [15], [16], [17], [18], [27], [19], [12]) and/or are restricted to a low subject number (e.g. < 13: [13], [15], < 25: [16], [12], the latter for OASIS and IBSR subsets). This is supposed to be problematic wrt. generalisation, but even inter- and intra rater variability of manual segmentations has only been investigated between two raters and 23 subjects [28]. Few studies / datasets exist relying on at most two manual raters: A possibility how to cope with this problem is exhibited by the Radiology Research Database[1] [29], which provides ground truth created by one rater and verified by two other raters. However, of course this approach involves laborious synchronisation among three raters and eventually, the verifying raters are already biased by the ground truth of the first rater. The approach in [30] uses 2 raters in a complicated and randomised collaborative protocol to avoid such a bias. Another approach chosen is to document inter-rater agreement of the rater employed in the study to another rater using a different (usually smaller) dataset [14], [16], which has the obvious problem of questionable generalisation of the results. In case multiple rater results are available per MRI volume, the generation of a reliable fused ground truth can be accomplished using the STAPLE technique [31] or any other technique as used in MAS label fusion like (locally weighted) majority voting [22]. While such techniques have of course been used in hippocampus MAS (e.g. [20]), to the authors knowledge this has not been applied so far to generate a single ground truth from multiple rater segmentations of the same MRI volume due to the non-availability of corresponding multiple ground truth hippocampal segmentations.

Another important issue for research efforts focusing on medical / clinical questions related to segmented hippocampi

[1] Available at http://www.nitrc.org/projects/hippseg_2011

but not on segmentation techniques itself is that without proper background and significant experience, a re-implementation of proposed techniques is far from being trivial and usually requires several man-years of programming effort. Therefore, especially for research groups "only" interested in segmentation results for further analysis, available (preferably cost-free) out-of-the-box segmentation software without the need for extensive optimisation and adaption is a highly attractive (if not the only) option. Screening corresponding software repositories and websites does not bring too many results – e.g. providing a web-service for single volumes to be segmented like at https://hipposeg.cs.ucl.ac.uk is highly valuable for educational purposes, but actual employment in a study requires batch-processing capabilities. So far, the authors were not able to spot easy-to-use publicly available MAS software for hippocampus segmentation.

In this paper, we assess variability in various segmentation reliability and accuracy aspects when using three cost-free and pre-compiled out-of-the-box hippocampus segmentation software packages. Section 2 describes the experimental setup in terms of dataset used, employed segmentation techniques, and evaluation methodology. In Section 3, we present segmentation results in terms of volume variability, shape accuracy variability among techniques and subjects, as well as ground truth variability among three manual raters and apply a simple majority voting label fusion technique to profit from this available ground truth at low cost.

II. Methodology

A. Data and Ground Truth

In this work we use a data set that has been acquired at the Salzburg Paracelsus Medical University and consists of 56 T1-weighted MRI volumes, including patients with mild cognitive impairment (MCI, 20 subjects), with temporal lobe epilepsy (TLE, 17 subjects), and a healthy control group (CG, 19 subjects). One of the subjects has been removed from the set since one of the programs evaluated consistently failed to produce a segmentation result for that subject. Hence, the final set consists of 55 volumes only, containing 27 males (18-76 years, mean age 53±19 years) and 28 females (23-71 years, mean age 54±14 years). In this set we have manual segmentations from 3 experienced raters (one senior neurosurgeon and two junior neuroscientists supervised by a senior neuroradiologist) for 9 identical subjects. This reduced set consists of 4 males (20-59 years, mean age 42 ± 17 years) and 5 females (28-49 years, mean age 47±17 years), all but one with diagnosed TLE and the remaining woman with MCI. For 39 subjects we have one manual segmentation from a single rater (also part of the above rater group). In this subset we have 17 males (18-74 years, mean age 46±20 years) and 22 females (23-71 years, mean age 51±15 years). This subset includes 4 MCI, 17 TLE, and 18 CG subjects, respectively.

Manual segmentations have been performed on a Wacom Cintiq 22HD graphic tablet device (resolution 1920x1200) using a DTK-2200 pen and employing the 32-bit 3DSlicer software for Windows (v. 4.2.2-1 r21513) to delineate hippocampus voxels for each slice separately. The raters independently used consensus on anatomical landmarks/boarders of the hippocampus based on Henry Duvemoy's hippocampal anatomy [32]. The procedure used was to depict the hippocampal outline in the view of all planes in the following order: sagittal – coronal – axial with subsequent cross line control through all planes.

B. Software Packages

We initially intended to use four different software packages, each relying on a different algorithmic principle, in the context of an automated segmentation of hippocampi. However, **AutoSeg**[2], although found to be usable to segment e.g. the Radiology Research Database [29], [33], was not applicable in this study due to repeated and enduring failures during the skull stripping process. In contrast to most of the algorithms presented in literature, e.g. [12], all these software packages are already pre-compiled and available for free.

FreeSurfer[3] is a popular set of tools which allow an automated labelling of subcortical structures in the brain. Such a subcortical labelling is obtained by using the volume-based stream which consists of five stages [10]. The result is a label volume, containing labels for various different subcortical structures (e.g. hippocampus, amygdala, and cerebellum). FreeSurfer is a highly popular tool to assess clinical hypotheses [30], [27], [19], [17], [15] or to compare to newly proposed segmentation techniques (e.g. [13], [16], [12]).

AHEAD (Automatic Hippocampal Estimator using Atlas-based Delineation[4]) is specifically targeted at an automated segmentation of hippocampi [11].

After an initial rigid registration step, a deformable registration is carried out using the Symmetric Normalisation algorithm. From the result of these steps, the volume is normalised to the atlas. The hippocampus segmentation from the atlas is then warped back to the input volume. Based on multiple atlases and a statistical learning method, the final segmentation is obtained.

Although **BrainParser**[5] is usually able to label various different subcortical structures, we use a version of BrainParser which is specifically tailored to hippocampus segmentation. After re-orienting the input volume to the coordinate system of the included, pre-trained atlas, skull stripping is performed. This is followed by computing an affine transform between the input volume and the reference brain volume. Then a deformable registration between the input and the reference volume is carried out. Then, according to the trained atlas, the input volume is labelled.

C. Segmentation Quality Assessment Metrics

To allow inter-rater comparisons of ground truth segmentations as well as assessment of the quality of the automated hippocampus segmentation methods, metrics are needed.

[2] v. 2.9, available at http://www.nitrc.org/projects/

[3] v. 51.0, available at http://surfer.nmr.mgh.harvard.edu

[4] v. 1.0, available at http://www.nitrc.org/projects/

[5] available at http://www.nitrc.org/projects/

In the following the automated segmentation is denoted by S, the ground truth segmentation is called G, and $v(\cdot)$ is a volume operator which computes the volume of a voxel volume with respect to the actual dimensions of a voxel.

- **Similarity index (SI)**
  The similarity index (also known as the Dice coefficient) is a quite frequently used measure to assess the similarity between two sets of voxels.
  $$SI(G,S) = \frac{2v(G \cap S)}{v(G) + v(S)} \tag{1}$$
- **Symmetric Hausdorff distance (SHD)**
  This metric is based on the actual structure of a voxel volume. It is defined as
  $$SHD(G,S) = \max(HD(G,S), HD(S,G)) \tag{2}$$
  where
  $$HD(X,Y) = \max_{x \in X}(\min_{y \in Y} d(x,y)). \tag{3}$$
  is the non-symmetric Hausdorff distance, x and y are vectors in $\mathbb{R}^3$ and $d(\cdot,\cdot)$ denotes the Euclidean distance between two vectors.

While low values in $[0,1]$ correspond to little similarity / quality for SI, the SHD produces large values (differences) between dissimilar segmentations.

III. Experimental Segmentation Results

The following results are always based on both hippocampi simultaneously. That is, we do not present results for the left and right hippocampus separately but treat both hippocampi from each scan as one segmentation object.

A. Volume Variability

Figure 1 provides an overview of the segmentation outcomes of the different program packages and three raters (thus 9 subjects are covered). In particular, this figure shows the volumes (in mm^3) for the hippocampi segmented from each subject, ignoring shape variations of course.

This figure shows that FreeSurfer segmentations have much higher volumes as compared to the raters, while AHEAD and BrainParser yield volumes comparable to those of the manual raters (corresponding results are also confirmed for all subjects, not displayed). FreeSurfer tends to over-segmentations in general, which has also been shown already in other studies [27], [19], thus demonstrating that our setup produces reasonable results in accordance with previous literature.

We also observe from this figure that rater variability is in the same order of magnitude as the differences between AHEAD and BrainParser, at least for these volume results.

B. Automated Segmentation Variability

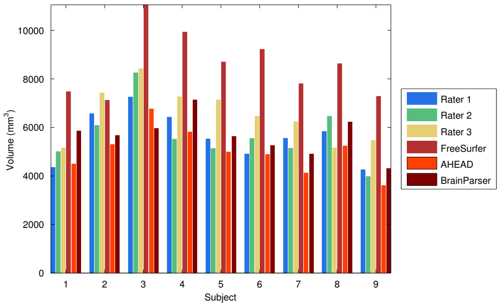


Figure 1. Comparison between the volumes for the automated segmentations and the volumes for the manual segmentations.

Table I shows the metric results for the similarity index and the symmetric Hausdorff distance, respectively, when comparing the segmentation outcomes among the automated segmentation programs, thus now reflecting shape similarity as well (mean scores along with the respective standard deviation over all subjects). The results are shown for the complete set of subjects (55) and for the subset for which three manual segmentations from human raters were available (9).

The results clearly show that the similarity between automated segmentations is in general highest when comparing BrainParser and AHEAD. As soon as FreeSurfer is compared to one of the other two software packages, the similarity gets lower (which is not surprising given the volume results in the previous section).

The rather high standard deviations in case of comparisons with BrainParser are the result of the incorrect segmentation of BrainParser for subject 38 (in fact, although a voxel volume has been returned by BrainParser, the intersection between this volume and the segmentations of the other two software packages is empty).

Table I
Summary of the results from the program comparisons.

	All subjects (55)		Reduced set (9)	
	SI	**SHD**	**SI**	**SHD**
AHEAD/FreeSurfer	0.67±0.05	9.27±1.87	0.65±0.04	9.90±1.14
AHEAD/BrainParser	0.71±0.13	7.18±8.36	0.75±0.03	6.03±2.11
BrainParser/FreeSurfer	0.67±0.13	10.25±8.22	0.68±0.05	10.87±3.93

C. Ground Truth Variability

Figure 2 shows the results for the symmetric Hausdorff distance, when comparing the manual segmentations of the human raters. The results in this figure of course show inter-rater variability.

Rater 3 causes rather huge distances in case of 3 subjects (no. 5,8,9). This suggests that there is a higher disagreement between rater 3 and the other raters then it is the case for the third rater pair. However, for subjects no. 2 and 4, respectively, the involvement of rater 2 causes high differences and for subject no. 7 only the difference between rater 1 and 2 stands out.

A summary of the results is given in Table II (mean and standard deviations over all subjects). From the results in this table we also notice that the rater pairs involving rater 3 have indeed a lower level of agreement. When comparing these scores with the ones for the program packages for the same subjects (reduced set results as given in Table I), we notice that the inter-rater agreements are at about the same level as they are in case of the comparisons between the automated segmentations. Results involving rater 3 are all in

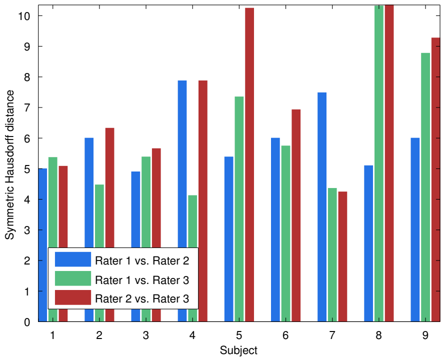


Figure 2. Comparison between the human raters (pairwise symmetric Hausdorff distance).

Table II
SUMMARY OF THE RESULTS FROM THE INTER-RATER COMPARISONS (9 SUBJECTS).

	SI	SHD
Rater 1 vs. Rater 2	0.79±0.04	5.97±1.01
Rater 1 vs. Rater 3	0.74±0.10	6.21±2.03
Rater 2 vs. Rater 3	0.74±0.10	7.33±2.12
Overall	0.76±0.07	6.50±1.30

all slightly worse than the best inter-software agreement, rater 1 and 2 agreement is slightly better as the best inter-program agreement. If we just compare the inter-rater agreements with the inter-program agreements of BrainParser and AHEAD, the levels of agreement between the raters and the programs are rather similar.

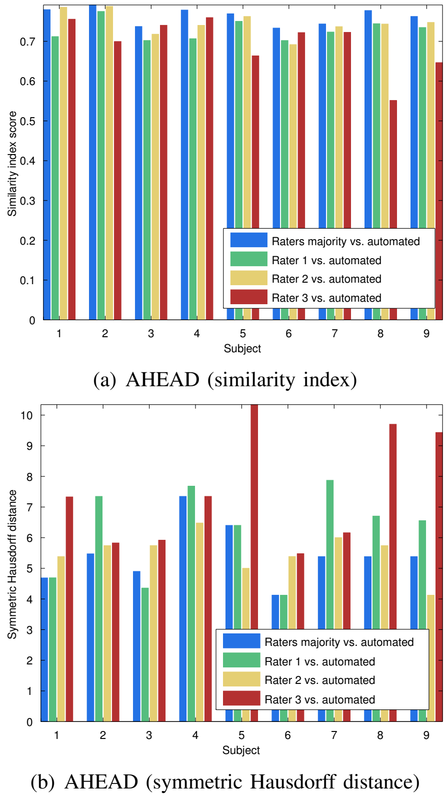
This result questions the usefulness of ground truth to rate software segmentation results for our purpose in general. Seeing that result variability is on comparable level for human raters (ground truth) and software (to be assessed) one might just think to pick an arbitrary software package.

D. Segmentation Techniques and Subject Variability

Results of the previous section obviously suggest not to rely on ground truth of single raters. A possibility how to cope with this problem is exhibited by the Radiology Research Database[6] [29], which provides ground truth created by one rater (and verified by two other raters). However, of course this approach involves tedious synchronisation among three raters and of course, the verifying raters are already biased by the ground truth of the first rater. In this section we follow a different strategy by using voxel-based majority voting among the segmentations of the three raters (a voxel is active in the fused volume if at least two raters marked that voxel as belonging to a hippocampus).

Figures 3 – 5 show example scores from the comparisons between the raters and the three automated segmentation programs used. Each plot shows the agreement between the single raters and each program as well as the agreement with the fused ground truth.

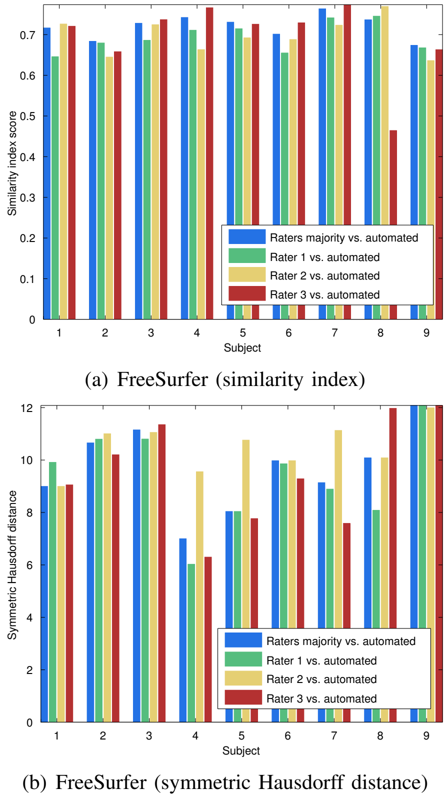
We see that in most cases, the fused rater segmentations lead to higher agreements with the program packages as



(a) AHEAD (similarity index)

(b) AHEAD (symmetric Hausdorff distance)

Figure 3. Comparison between the human raters and AHEAD segmentation.

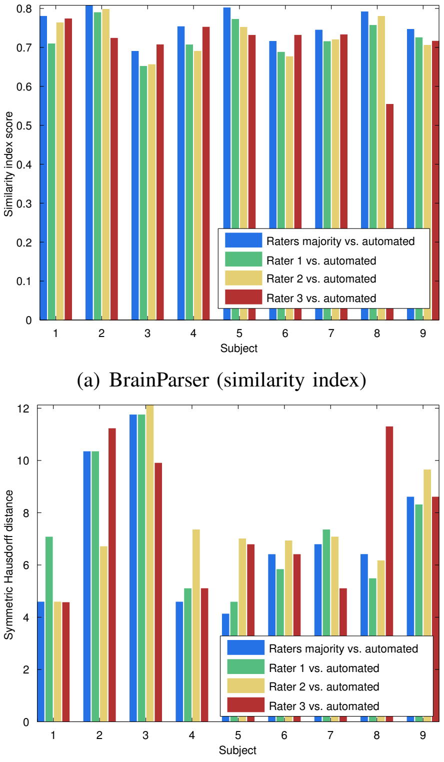


(a) FreeSurfer (similarity index)

(b) FreeSurfer (symmetric Hausdorff distance)

Figure 4. Comparison between the human raters and FreeSurfer segmentation.

[6] Available at http://www.nitrc.org/projects/hippseg_2011



(a) BrainParser (similarity index)

(b) BrainParser (symmetric Hausdorff distance)

Figure 5. Comparison between the human raters and BrainParser segmentation.

compared to the single raters. In any case, extreme agreements and disagreements vanish making the assessment significantly more reliable.

Table III provides a summary of all corresponding results (mean over all subjects and raters – first three lines – and mean over all majority voting results over all subjects for the remaining lines). From this table we again notice two things: First, in case of FreeSurfer the similarities are lower as compared to the other programs, while AHEAD provides the best results wrt. both, best mean quality and lowest standard deviation. And, second, using the fused rater segmentations enhances the scores in terms of mean quality but consistently leads to higher standard deviations, exhibiting the differences of the techniques even clearer.

Table III
SUMMARY OF THE RESULTS FROM THE COMPARISONS BETWEEN THE RATERS AND THE PROGRAMS (9 SUBJECTS).

	SI	SHD
FreeSurfer vs. Raters	0.69±0.03	9.80±1.38
AHEAD vs. Raters	0.72±0.02	6.40±0.86
BrainParser vs. Raters	0.72±0.03	7.49±1.96
FreeSurfer vs. Majority	0.72±0.03	9.68±1.57
AHEAD vs. Majority	0.76±0.02	5.45±0.95
BrainParser vs. Majority	0.76±0.04	7.06±2.67

Since we had forty manual segmentation in case of rater 2, we also did a comparison of those manual segmentations and the software packages.

In case of 4 subjects (no. 10, 22, 31, and 34) the results are rather poor in case of the rater comparison against BrainParser. While for subject 34 there is no BrainParser segmentation available (BrainParser was not able to register the scan against the atlas used, hence, no segmentation result is available), BrainParser yielded erroneous segmentations in case of the other subjects (e.g. due to a failed skull stripping). The results for subject 34 are therefore excluded from the summary given in the table below.

The general picture as shown in Table IV is that the agreements between rater 2 and the automated segmentation programs are rather high. Again, when compared against FreeSurfer, the similarities between the segmentations get lower due to the oversegmentation. This table also shows that the agreement between rater 2 and BrainParser is a bit lower as compared to AHEAD. This is caused by the problematic subjects listed above.

Table IV
SUMMARY OF THE RESULTS FROM THE COMPARISONS BETWEEN RATER 2 AND THE PROGRAMS (39 SUBJECTS).

	SI	SHD
FreeSurfer vs rater 2	0.71±0.06	10.04±2.37
AHEAD vs rater 2	0.73±0.05	6.48±2.04
BrainParser vs rater 2	0.69±0.16	9.21±9.94

Thus, overall, AHEAD again provides the best results in terms of mean quality and lowest standard deviation, confirming the results gathered from the reduced dataset with 9 subjects.

One important point to consider when looking at the summarised results is that the segmentation accuracy of a rater may vary considerably from subject to subject. This of course also applies to the different program packages. To illustrate this, we selected two subsets of 9 subjects from the 39 subjects available and compared the mean agreement between rater 2 and the programs. One subset consists of subjects for which there is only a rather low agreement between the rater and the programs, whereas the second set consists of subjects with a high agreement (the sets have been created for each program/metric combination separately, see Table V for results).

Table V
SUMMARY OF THE RESULTS FROM THE COMPARISONS BETWEEN RATER 2 AND THE PROGRAMS (ALWAYS FOR 9 SUBJECTS).

	Low agreement		High agreement	
	SI	SHD	SI	SHD
FreeSurfer/rater	0.62±0.07	13.17±2.42	0.75±0.00	8.61±0.00
AHEAD/rater	0.67±0.05	9.13±2.90	0.77±0.00	5.38±0.00
BrainParser/rater	0.49±0.24	18.44±18.46	0.77±0.00	6.00±0.00

From this table we immediately see that there is indeed a huge impact on the scores, if the set of segmentations at hand contains either inaccurate segmentations by the manual rater or inaccurate segmentations by the programs. In case of the low agreement set, BrainParser is now clearly the worst technique, while in the case of the high agreement set, it is no longer possible to clearly identify the best performing technique.

IV. Conclusion

We have found ground truth variability and automatic segmentation results variability on a comparable level, which makes the segmentation accuracy assessment based on a single-rater ground truth hardly reliable. This is especially surprising since most of the subjects involved suffer from TLE and MCI, and automated segmentation techniques rely on atlases of healthy subjects. Thus, a clear superiority of manual segmentations would have been expected.

Using a majority voting based label fusion ground truth, we were able to identify AHEAD as the most reliable automated segmentation tool considered. Having observed the importance of subject variability, results still need to be strengthened with a larger set of manual segmentations in future work.

Acknowledgments

This has been funded by the Austrian Science Fund (FWF) under Project No. KLI 00012.

References

[1] M. Fotuhi, D. Do, and C. Jack, "Modifiable factors that alter the size of the hippocampus with ageing," *Nat Rev Neurol*, vol. 8, pp. 189–202, 2012.

[2] K. Malmgren and M. Thom, "Hippocampal sclerosis - origins and imaging," *Epilepsia*, vol. 53, pp. 19–33, 2012.

[3] H. Eichenbaum, "What H.M. taught us," *J Cogn Neurosci*, vol. 25, pp. 14–21, 2013.

[4] C. Butler and A. Zeman, "Recent insights into the impairment of memory in epilepsy: transient epileptic amnesia, accelerated long-term forgetting and remote memory impairment," *Brain*, vol. 131, pp. 2243–2263, 2008.

[5] B. Winblad, K. Palmer, M. Kivipelto, V. Jelic, L. Fratiglioni, L. Wahlund *et al.*, "Mild cognitive impairment - beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment," *J Intern Med*, vol. 256, pp. 240–246, 2004.

[6] E. Korf, L. Wahlund, P. Visser, and P. Scheltens, "Medial temporal lobe atrophy on MRI predicts dementia in patients with mild cognitive impairment," *Neurology*, vol. 63, no. 1, pp. 94–100, 2004.

[7] C. Geroldi, R. Rossi, C. Calvagna, C. Testa, L. Bresciani, G. Binetti, O. zanetti, and G. Frisoni, "Medial temporal atrophy but not memory deficits predict progression to dementia in patients with mild cognitive impairment," *J Neurol Neurosurg Psychiatry*, vol. 77, pp. 1219–1222, 2006.

[8] P. Borghesani, S. DeMers, V. Manchanda, S. Pruthi, D. Lewis, and S. Borson, "Neuroimaging in the clinical diagnosis of dementia: Observations from a memory disorders clinic," *JAGS*, vol. 58, pp. 1453–1458, 2010.

[9] S. Teipel, M. Grothe, S. Lista, N. Toschi, F. Garaci, and H. Hampel, "Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease," *Med Clin North Am*, vol. 97, pp. 399–424, 2013.

[10] B. Fischl, A. van der Kouwe, C. Destrieux, E. Halgren, F. Ségonne, D. H. Salat, E. Busa, L. J. Seidman, J. Goldstein, D. Kennedy, V. Caviness, N. Makris, B. Rosen, and A. M. Dale, "Automatically parcellating the human cerebral cortex," *Cerebral Cortex*, vol. 14, no. 1, pp. 11–22, 2004.

[11] J. W. Suh, H. Wang, S. Das, B. Avants, and P. A. Yushkevich, "Automatic segmentation of the hippocampus in t1-weighted mri with multi-atlas label fusion using open source software: Evaluation in 1.5 and 3.0t adni mri," in *Proceedings of the International Society for Magnetic Resonance in Medicine conference (ISMRM'11)*, 2011.

[12] D. Zarpalas, P. Gkontra, P. Daras, and N. Maglaveras, "Accurate and fully automatic hippocampus segmentation using subject-specific 3d optimal local maps into a hybrid active contour model," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 2, pp. 1–16, 2014.

[13] H. Pardoe, G. Pell, D. Abbott, and G. Jackson, "Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation?" *Epilepsia*, vol. 50, no. 12, pp. 2586–2592, 2009.

[14] G. Winston, M. Cardoso *et al.*, "Automated hippocampal segmentation in patients with epilepsy: Available free online," *Epilepsia*, vol. 54, no. 12, pp. 2166–2173, 2013.

[15] J. Kim, D. Choi *et al.*, "Evaluation of hippocampal volume based on various inversion time in normal adults by manual tracing and automated segmentation methods," *Investig Magn Reson Imaging*, vol. 19, no. 2, pp. 67–75, 2015.

[16] R. Morey, C. Petty *et al.*, "A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes," *Neoroimage*, vol. 45, no. 3, pp. 855–866, 2009.

[17] H. Kim, M. Chupin, O. Colliot *et al.*, "Automatic hippocampal segmentation in temporal lobe epilepsy: Impact of developmental abnormalities," *Neoroimage*, vol. 59, pp. 3178–3186, 2012.

[18] S. Babakchanian, N. Chew, A. Green *et al.*, "Automated and manual hippocampal segmentation techniques: A comparison of results and reproducibility," *Neorology*, vol. 80, p. P06.053, 2013.

[19] W. S. Tae, S. S. Kim, K. U. Lee, E.-C. Nam, and K. W. Kim, "Validation of hippocampal volumes measured using a manual method and two automated methods (freesurfer and ibaspm) in chronic major depressive disorder," *Neuroradiology*, vol. 50, pp. 569–581, 2009.

[20] J. Cardoso, K. Leung *et al.*, "STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation," *Medical Image Analysis*, vol. 17, no. 6, pp. 671–684, June 2013.

[21] K. Leung, J. Barnes *et al.*, "Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease," *NeuroImage*, vol. 51, no. 4, pp. 1345–1359, 2013.

[22] J. Iglesias and M. Sabancu, "Multi-atlas segmentation of biomedical images: a survey," *Medical Image Analysis*, vol. 24, pp. 205–219, 2015.

[23] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady, "Evaluating segmentation error without ground truth," in *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'12)*, ser. Lecture Notes in Computer Science, vol. 7510, 2012, pp. 528–536.

[24] J. Bogovic *et al.*, "Approaching experts results using a hierarchical cerebellum parcellation protocol for multipe inexpert human raters," *NeiroImage*, vol. 64, pp. 616–629, 2013.

[25] L. Maier-Hein, D. Kondermann *et al.*, "Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences," *Int. J. Computer Assisted Radiology and Surgery*, vol. 10, no. 8, pp. 1201–1212, 2015.

[26] R. Kwitt, S. Hegenbart, N. Rasiwasia, A. Vécsei, and A. Uhl, "Do we need annotation experts? a case study in celiac disease classification," in *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'14)*, ser. Lecture Notes in Computer Science, vol. 8674, September 2014, pp. 454–461.

[27] N. Cherbuin, K. J. Anstey1, C. Réglade-Meslin, and P. S. Sachdev, "In vivo hippocampal measurement and memory: A comparison of manual tracing and automated segmentation in a large community-based sample," *PLoS ONE*, vol. 4, pp. 1–10, 2009.

[28] E. Achten, K. Deblaere, C. D. Wagter *et al.*, "Intra- and interobserver variability of mri-based volume measurements of the hippocampus and amygdala using the manual ray-tracing method," *Neuroradiology*, vol. 40, no. 9, pp. 558–566, 1998.

[29] K. Jafari-Khouzani, K. Elisevich, S. Patel, and H. Soltanian-Zadeh, "Database of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques," *Neuroinformatics*, vol. 9, no. 4, pp. 335–346, 2011.

[30] E. Wenger, J. Martensson, H. Hoack *et al.*, "Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains," *Epilepsia*, vol. 35, no. 8, pp. 4236–4248, 2014.

[31] S. Warfield, K. Zou, and W. Wells, "Simulataneous truth and performance level estimation (STAPLE): an algortihm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[32] R. Kuzniecky and G. Jackson, *Magnetic resonance in epilepsy*. New York: Raven Press, 1995.

[33] M. Gschwandtner, Y. Höller, M. Liedlgruber, E. Trinka, and A. Uhl, "Assessing out-of-the-box software for automated hippocampus segmentation," in *Proceedings of Bildverarbeitung für die Medizin 2016 (BVM'16)*, ser. Springer Informatik Aktuell, pp. 212–217, Mar. 2016.