Psychiatry Research: Neuroimaging 280 (2018) 39–47

Contents lists available at ScienceDirect

Psychiatry Research: Neuroimaging

journal homepage: www.elsevier.com/locate/psychresns

Can measuring hippocampal atrophy with a fully automatic method be substantially less noisy than manual segmentation over both 1 and 3 years?

Keith S. Cover[a,*], Ronald A. van Schijndel[a], Paolo Bosco[b], Soheil Damangir[c], Alberto Redolfi[d], Alzheimer's Disease Neuroimaging Initiative[1]

[a] Amsterdam University Medical Center, Amsterdam, The Netherlands
[b] National Institute for Nuclear Physics, Pisa, Italy
[c] Karolinska Institutet, Stockholm, Sweden
[d] IRCCS San Giovanni di Dio Fatebenefratelli, Brescia, Italy

ARTICLE INFO

Keywords:
Hippocampus
Atrophy
Manual segmentation
Automatic segmentation
Magnetic resonance imaging
Alzheimer's disease
Binomial

ABSTRACT

To quantify the "segmentation noise" of several widely used fully automatic methods for measuring longitudinal hippocampal atrophy in Alzheimer's disease and compare the results to the segmentation noise of manual segmentation over both 1 and 3 years. The segmentation noise of 5 longitudinal hippocampal atrophy measurement methods was quantified, including checking its Gaussianity, using 264 subjects from the ADNI1 back-to-back (BTB) data set over both 1 year and 3 year intervals. The segmentation methods were FreeSurfer 5.3.0 both cross sectional and longitudinal, FreeSurfer 6.0.0 longitudinal, MAPS-HBSI and FSL/FIRST 5.0.8. The BTB manual segmentation of 75 ADNI subjects from a previous study provided the manual distributions for comparison. All methods, including the manual segmentation, violated the Gaussianity assumption. Two methods, FreeSurfer 6.0.0 and MAPS-HBSI, had a segmentation noise substantially less than a surrogate for manual segmentation. FreeSurfer 5.3.0 longitudinal was confirmed as a surrogate for manual segmentation. The violation of the Gaussian assumption by the segmentation methods assessed, including manual, suggests results of previous studies that assumed Gaussian statistics without confirmation may need review. Fully automatic FreeSurfer 6.0.0 and MAPS-HBSI both have lower segmentation noise than manual requiring less than two thirds of the subjects to detect the same treatment effect.

1. Introduction

Hippocampal atrophy is the amount of shrinkage of the hippocampus from one time point to the next. It can be measured with noninvasive MRI and is a widely validated surrogate outcome for Alzheimer's disease (AD) trials (Frisoni et al., 2010). It has been shown to be one of the first observable characteristics of AD (Bobinski et al., 1996). It also accelerates before the translation to clinical dementia (Jack et al., 2011) as part of the AD pathology cascade (Jack et al., 2010). Analysis of the images from the ADNI1 study found the median annualized atrophy rates were 1.5% (healthy controls (HC)), 2.4% (mildly cognitively impaired (MCI)) and 5.1% (AD) (Cover et al., 2016).

Many software methods are available to fully automatically measure hippocampal atrophy directly (longitudinal measurement) or indirectly by measuring the change in volume between two time points (cross sectional measurement). Segmentation methods include FreeSurfer (Fischl et al., 2012), FSL/FIRST (Patenaude et al., 2011) and MAPS-HBSI (Leung et al., 2010). Several fully automatic segmentation methods also have government approval for clinical use including NeuroReader (Ahdidan et al., 2015; NeuroReader, 2016), LEAP (Woltz et al., 2014; LEAP, 2016) and NeuroQuant (Ochs et al., 2015; NeuroQuant, 2016).

Correctly assessing the performance of fully automatic segmentation methods is particularly challenging when the methods perform better than the gold standard of manual segmentation. Recently, we reported that a fully automatic segmentation method (MAPS-HBSI) (Leung et al.,

* Corresponding author.
E-mail address: keith@kscover.ca (K.S. Cover).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.pscychresns.2018.06.011
Received 19 February 2018; Received in revised form 26 June 2018; Accepted 27 June 2018
Available online 09 August 2018
0925-4927/

2010) had substantially lower segmentation noise than manual segmentation (Cover et al., 2016).

Almost all studies assessing the segmentation noise of the hippocampus - and other structures such as the whole brain or cortical thickness - have used parametric statistics which assume Gaussian distributions - such as the mean, standard deviation and interclass correlation coefficient. Gaussian distributions are also referred to as normal distributions. There are only a few exceptions (Smith et al., 2007; Cover et al., 2011; Mulder et al., 2014; Cover et al., 2016; Opfer et al., 2016). The validity of the parametric statistics rests on the segmentation noise having a Gaussian distribution. While the segmentation noise of the whole brain has been shown to violate the Gaussian assumption (Cover et al., 2011), no study in the literature has checked whether any of the segmentation methods measuring hippocampal atrophy has a Gaussian noise distribution. Rarely has the potential impact of non-Gaussian distributions on parametric statistical calculations such as sample size been considered.

The current study focuses on assessing the segmentation noise - as measured by the back-to-back (BTB) reproducibility - of hippocampal atrophy measuring methods that have lower segmentation noise than that of the manual method. The segmentation noise for all methods is analyzed with both parametric and robust statistical methods. Also, the Gaussianity of the segmentation noise distributions is checked to determine if robust statistics are required. In addition, the segmentation noise of FreeSurfer 5.3.0 longitudinal is compared to the segmentation noise of manual measurements to confirm FreeSurfer 5.3.0 longitudinal is a suitable surrogate for the noise of manual segmentation. Finally, the segmentation noise over 1 and 3 years for all methods is compared to the surrogate for manual segmentation noise.

2. Methods

2.1. Dataset

The ADNI1 data set is widely used in studies of the reproducibility of structural measures including the hippocampus (Cover et al., 2011; Mulder et al., 2014; Ochs et al., 2015; Ahdidan et al., 2015; Chincarini et al., 2016; Cover et al., 2016). The 1.5T T1-weighted MRI scans were selected from the ADNI database and downloaded in their original unprocessed DICOM format. A total of 264 subjects are selected that had two BTB scans at baseline, 1 year and 3 years for a total of $6 \times 264 = 1{,}584$ image volumes. Supplemental table S1 has a complete listing of the subjects used including exact identification of the image volume. The 264 subjects in the current study are a subset of the 562 ADNI1 BTB 1.5T subjects in a previous study (Cover et al., 2016). Only 264 of the 562 subjects also had BTB scans at 3 years in ADNI1 therefore only 264 subjects are used in the current study.

The ADNI1 study acquired the MRI sequence twice during each patient visit. The subjects did not leave the MRI between MPRAGEs and often the second MPRAGE started within a few second of the completion of the first. While the images from only one MPRAGE sequence at each patient visit are needed to calculate the hippocampal atrophy, the second MPRAGE provides excellent data to make noise measurements. The two MPRAGE sequences are referred to as BTB, rather than scan-rescan, because they were acquired without the patient leaving the MRI scanner. The topic of the current paper is the noise of the segmentation methods. It is a reasonable assumption that all the segmentations methods in the current paper are relatively accurate as they are widely used. Thus, the accuracy of the segmentation methods is beyond the scope of the current paper.

The 264 subjects in the current study contained 120 healthy controls (HC), 143 mildly cognitively impaired (MCI) and 1 probable AD as classified by the ADNI1 study The low number of probable AD subjects is likely due to the higher probability of probable AD subjects dropping from the study over the first 3 years. Table 1 provides descriptive statistics of the 264 subjects.

Table 1
Descriptive statistics of the 264 subjects from the ADNI1 included in the current study. The interquartiles are in brackets.

Cohort	Status	Sample size	M/F	Age (Baseline)
3 year fully automatic	HC	120	66/54	75.0 (72.0, 78.5)
	MCI	143	100/43	74.1 (70.6, 80,5)
	AD	1	1/0	78.4
	Combined	264	167/97	74.4 (71.5, 79.5)
1 year manual	HC	19	11/8	76.5 (72.1, 79.6)
	MCI	38	25/13	73.7 (70.7, 77.9)
	AD	18	7/11	74.1 (69.4, 78.4)
	Combined	75	43/32	74.1 (70.7, 77.9)

As no manual segmentation was performed as part of the current study, 75 ADNI1 BTB 1.5T subjects used in prior studies (Mulder et al., 2014; Cover et al., 2016) were also included in the current study to provide some statistics on the performance of manual segmentation. While the 75 subjects are also a subset of the 562 subjects used in a previous study (Cover et al., 2016), as are the 264 subjects mentioned above, only 40 of the subjects were common to both.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI study was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

2.2. Statistical Analysis

A detailed description of the statistical analysis for BTB atrophy measurement has been presented previously (Cover et al., 2016). Fig. 1 of the current paper provides the steps to calculate the BTB differences over 1 year and 3 years. Additional details of the calculations follow.

The amount of atrophy - as measured by the percentage volume change (PVC) - from baseline (V_A) to year 1 or year 3 (V_B) was calculated by the equation $100*(V_B\text{-}V_A)/V_A$. For each subject there were 8 PVCs - 2 for the left and right hippocampus, 2 for the 1 year and 3 years intervals and 2 for the BTB acquisition. The BTB differences were calculated by subtracting the PVCs of the first acquired image volume of a subject visit from that of the second. Consequently, there were 4 BTB differences for each subject - one each for the left and right hippocampus and one each for the 1 year and 3 year intervals. As a result, there were 4 BTB difference distributions for each method.

A variety of statistics were calculated for each BTB difference distribution. All statistics were calculated from the absolute values of the BTB differences. The statistics included the maximum, minimum, median (MDBTBD), mean (MNBTBD) and standard deviation (SDBTBD). The value of the mean subtracted off before calculating the standard deviation was assumed to be zero. The number of BTB differences in each distribution is also listed so the number of times each method failed to yield a BTB difference can be determined.

Three different statistical tests were used to test the Gaussianity of the BTB difference distributions. Two of the tests, the Anderson-Darling and the Shapiro-Wilk tests, tested general properties of Gaussianity. The third test was tailored to whether the distributions had too many outliers for a Gaussian distribution.

The tailored test is based on the ratio of SDBTBD and MDBTBD, two measures of the spread of BTB distributions used in the literature. For an ideal Gaussian distribution the ratio of SDBTBD/MDBTBD is 1.3654. However, as the standard deviation of a distribution is more sensitive to the distribution's shoulders the ratio increases as the shoulders get larger. To calculate the *p*-value for a range of ratios, 10,000,000