Psychiatry Research: Neuroimaging 252 (2016) 26–35

Contents lists available at ScienceDirect

Psychiatry Research: Neuroimaging

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

Reproducibility of hippocampal atrophy rates measured with manual, FreeSurfer, AdaBoost, FSL/FIRST and the MAPS-HBSI methods in Alzheimer's disease

Keith S. Cover [a,*], Ronald A. van Schijndel [a], Adriaan Versteeg [a], Kelvin K. Leung [b], Emma R. Mulder [a], Remko A. Jong [a], Peter J. Visser [a], Alberto Redolfi [c], Jerome Revillard [d], Baptiste Grenier [d], David Manset [d], Soheil Damangir [e], Paolo Bosco [c], Hugo Vrenken [a], Bob W. van Dijk [a], Giovanni B. Frisoni [c,f], Frederik Barkhof [a], for the Alzheimer's Disease Neuroimaging Initiative, neuGRID [1]

[a] VU University Medical Center, Amsterdam, Netherlands
[b] UCL Institute of Neurology, London, United Kingdom
[c] IRCCS San Giovanni di Dio Fatebenefratelli, Italy
[d] MAAT, Archamps, France
[e] Karolinska Institutet, Sweden
[f] University Hospitals and University of Geneva, Switzerland

ARTICLE INFO

Article history:
Received 26 June 2015
Received in revised form
16 February 2016
Accepted 8 April 2016
Available online 11 May 2016

Keywords:
Hippocampus
Atrophy
Manual segmentation
Automatic segmentation
Magnetic resonance imaging
Mild cognitive impairment
Alzheimer disease
Boundary shift integral

ABSTRACT

The purpose of this study is to assess the reproducibility of hippocampal atrophy rate measurements of commonly used fully-automated algorithms in Alzheimer disease (AD). The reproducibility of hippocampal atrophy rate for FSL/FIRST, AdaBoost, FreeSurfer, MAPS independently and MAPS combined with the boundary shift integral (MAPS-HBSI) were calculated. Back-to-back (BTB) 3D T1-weighted MPRAGE MRI from the Alzheimer's Disease Neuroimaging Initiative (ADNI1) study at baseline and year one were used. Analysis on 3 groups of subjects was performed – 562 subjects at 1.5 T, a 75 subject group that also had manual segmentation and 111 subjects at 3 T. A simple and novel statistical test based on the binomial distribution was used that handled outlying data points robustly. Median hippocampal atrophy rates were −1.1%/year for healthy controls, −3.0%/year for mildly cognitively impaired and −5.1%/year for AD subjects. The best reproducibility was observed for MAPS-HBSI (1.3%), while the other methods tested had reproducibilities at least 50% higher at 1.5 T and 3 T which was statistically significant. For a clinical trial, MAPS-HBSI should require less than half the subjects of the other methods tested. All methods had good accuracy versus manual segmentation. The MAPS-HBSI method has substantially better reproducibility than the other methods considered.

1. Introduction

A feature of Alzheimer's disease (AD) (Jack et al., 1992, 1998; Wang et al., 2003; Evans et al., 2010; Frisoni et al., 2010; Drago et al., 2011) is increased hippocampal volume loss when compared to age matched healthy controls (HC). Mildly cognitive impairment (MCI) subjects typically have intermediate hippocampal volumes and rates of loss. Hippocampal atrophy rates have been proposed (Schott et al., 2010; Ard and Edland, 2011) or used (Wilkinson et al., 2012) as end points in clinical trials. Manual segmentation of hippocampi (Barnes et al., 2008; Boccardi et al., 2011) is often regarded as the "gold standard" for volume measurement – however this may take about 3 h per MRI scan (Mulder et al., 2014) and requires extensive training. The size of AD clinical trials (typically many hundreds of subjects) means that there is great interest in less labour-intensive methods; as a result several fully automated techniques have been developed and are increasingly used.

* Correspondence to: Physics and Medical Technology, VU University Medical Center, 1007 MB Amsterdam, Netherlands.
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.

http://dx.doi.org/10.1016/j.pscychresns.2016.04.006
0925-4927/

Manual measurements of hippocampal volume or atrophy rate are generally assumed to be more accurate than automated methods (Barnes et al., 2008; Boccardi et al., 2011) and are used for validation of the accuracy of automated techniques (Hsu et al., 2002; Tae et al., 2008; Morey et al., 2009; Pardoe et al., 2009; Dewey et al., 2010; Lehmann et al., 2010; Sanchez-Benavides et al., 2010; Doring et al., 2011; Kim et al., 2012; Iglesiasa et al., 2015). However, fully automatic methods have improved to the point where it has been suggested that they have similar accuracy when compared to manual measures and are more reproducible (Duchesne et al., 2002; Kennedy et al., 2009; Dewey et al., 2010; Doring et al., 2011). As a consequence, a number of comparisons of methods for measuring atrophy rates have been published (Kikinis et al., 1992; Fox and Freeborough, 1997; Rudick et al., 1999; Crum et al., 2001; Zhang et al., 2001; Smith et al., 2002; Barnes et al., 2004, 2007; van de Pol et al., 2007; Altmann et al., 2009; Barkhof et al., 2009; Shaw et al., 2009; Sluimer et al., 2009; Shen et al., 2010; Westman et al., 2011).

The ideal way to compare atrophy rate measurement methods would use perfectly accurate segmentations as a gold standard. The performance of each method could then be compared against the perfect segmentation over a set of subjects. By calculating the spread of the errors in each method – such as the standard deviation – the best performing methods could be determined. Perfectly accurate segmentations are not available, but we can obtain an indication of the spread of the errors in the methods - provided the methods are reasonably accurate - by repeating the measurements and determining their spread.

The goal of the current study was to compare the reproducibility of hippocampal atrophy rate of commonly-used automated measurement techniques, at both 1.5 T and 3 T, taking advantage of back-to-back (BTB) MPRAGE volumetric scans routinely acquired at each subject in the first Alzheimer's Disease Neuroimaging Initiative (ADNI1) study. We aimed to assess the most recent versions of FreeSurfer (Fischl et al., 2002, 2004; Reuter et al., 2012), FSL/FIRST (Patenaude et al., 2011), AdaBoost (Morra et al., 2009) and MAPS-HBSI (Leung et al., 2010).

The data set from the ADNI1 study (Mueller et al., 2005; Jack et al., 2008; Weiner et al., 2012) provides a singular opportunity to compare the reproducibilities of brain atrophy methods. While rarely mentioned in the literature, as part of ADNI1, two 3D T1 weighted MPRAGEs were acquired BTB during each subject visit - with the acquisition of the second MPRAGE usually starting within seconds of completion of the first (Cover et al., 2011). All ADNI1 subjects were asked to have a scan at 1.5 T with a subset of subjects also having 3T imaging. With 800 subjects acquired across 55 sites included in ADNI1, it provides a much larger BTB dataset than available for previous reproducibilities studies. In addition, the ADNI1 study put a great deal of effort into standardizing the acquisition of the MPRAGE sequences across the ADNI1 sites. Thus, ADNI1 provides an excellent dataset to test the reproducibility of the measurement of hippocampal atrophy rates and other structural segmentation methods.

For the hippocampus atrophy rates, the BTB reproducibility of manual segmentation at 1.5 T of hippocampi atrophy (Mulder et al., 2014) has been compared to FreeSurfer, and FSL/FIRST for a subset of N = 80 subjects of the ADNI1 dataset. Mulder et al. found the manual and automated segmentations had similar reproducibilities.

Although the ADNI1 study was performed primarily at 1.5 T, with research studies and trials in AD and other disorders shifting to 3 T acquisitions (de Jong et al., 2008; Watson et al., 2010) it was important to include in ADNI1 a sub-set of subjects who had 3 T BTB as well as 1.5 T BTB imaging. A direct comparison between 3 T and 1.5 T has only been performed for a cross sectional method (Keihaninejad et al., 2010) but without reproducibility measurements. Longitudinally, only the reproducibility of the FSL/Siena measure for whole brain atrophy has been compared at 1.5 T and 3 T (Cover et al., 2014).

In addition, for whole brain volume atrophy measures at 1.5 T (Popescu et al., 2012), subsets of the ADNI1 BTB dataset have been used to compare the reproducibility (Cover et al., 2011) of Siena and SienaX.

Here, we compared the reproducibility of 7 popular methods to determine hippocampal atrophy rates over 1 year. Such information is important to plan clinical trials in AD.

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 was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5 year public-private partnership. 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). Determination of sensitive and specific markers of very early AD.

Progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California–San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55–90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

2. Materials and methods

2.1. ADNI1 dataset

In ADNI1 two BTB MPRAGEs were acquired with identical acquisition parameters during each of the two subject visits – baseline and year one – without removing the subject from the scanner (Jack et al., 2008). Referred to as “original” MPRAGEs by ADNI1, for the current study the first acquired original MPRAGE is referred to as “M” and the second as “N”. ADNI selected one of M or N for additional processing and produced a third MPRAGE - referred to as “processed” by ADNI - for each subject visit. The processed MPRAGE is referred to as “P” in this study. The additional ADNI processing to generate P included B1 non-uniformity correction, intensity nonuniformity correction and gradient warp correction (Jack et al., 2008; Clarkson et al., 2009). Fig. 1 illustrates the relationship of the 6 MPRAGEs for each subject. While M and N provide information on reproducibility, P provides accuracy information used to ensure the atrophy rates of the methods are accurate enough that the reproducibilities are meaningful.

The M, N and P MPRAGEs used in the current study are exactly those downloaded from ADNI. According to ADNI, the M and N voxel values – which were 16 bit values - are unchanged from those generated by the MRI scanners. Only some of the meta data