NeuroImage: Clinical 11 (2016) 802–812

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity

Christopher G. Schwarz[a,*], Jeffrey L. Gunter[a,b], Heather J. Wiste[c], Scott A. Przybelski[c], Stephen D. Weigand[c], Chadwick P. Ward[a], Matthew L. Senjem[a,b], Prashanthi Vemuri[a], Melissa E. Murray[d], Dennis W. Dickson[d], Joseph E. Parisi[e], Kejal Kantarci[a], Michael W. Weiner[f], Ronald C. Petersen[g], Clifford R. Jack Jr [a], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]Department of Radiology, Mayo Clinic and Foundation, Rochester, MN, USA
[b]Department of Information Technology, Mayo Clinic and Foundation, Rochester, MN, USA
[c]Department of Health Sciences Research, Division of Biostatistics, Mayo Clinic and Foundation, Rochester, MN, USA
[d]Department of Neuroscience (Neuropathology), Mayo Clinic and Foundation, Jacksonville, FL, USA
[e]Department of Laboratory Medicine, Mayo Clinic and Foundation, Rochester, MN, USA
[f]Veterans Affairs, University of California, San Francisco, CA, USA
[g]Department of Neurology, Mayo Clinic and Foundation, Rochester, MN, USA

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form 29 April 2016
Accepted 27 May 2016
Available online 30 May 2016

ABSTRACT

Alzheimer's disease (AD) researchers commonly use MRI as a quantitative measure of disease severity. Historically, hippocampal volume has been favored. Recently, "AD signature" measurements of gray matter (GM) volumes or cortical thicknesses have gained attention. Here, we systematically evaluate multiple thickness- and volume-based candidate-methods side-by-side, built using the popular FreeSurfer, SPM, and ANTs packages, according to the following criteria: (a) ability to separate clinically normal individuals from those with AD; (b) (extent of) correlation with head size, a nuisance covariatel (c) reliability on repeated scans; and (d) correlation with Braak neurofibrillary tangle stage in a group with autopsy. We show that volume- and thickness-based measures generally perform similarly for separating clinically normal from AD populations, and in correlation with Braak neurofibrillary tangle stage at autopsy. Volume-based measures are generally more reliable than thickness measures. As expected, volume measures are highly correlated with head size, while thickness measures are generally not. Because approaches to statistically correcting volumes for head size vary and may be inadequate to deal with this underlying confound, and because our goal is to determine a measure which can be used to examine age and sex effects in a cohort across a large age range, we thus recommend thickness-based measures. Ultimately, based on these criteria and additional practical considerations of run-time and failure rates, we recommend an AD signature measure formed from a composite of thickness measurements in the entorhinal, fusiform, parahippocampal, mid-temporal, inferior-temporal, and angular gyrus ROIs using ANTs with input segmentations from SPM12.

1. Introduction

Accurate, reliable quantitative measures of disease severity from magnetic resonance imaging (MRI) have been a mainstay of research in the imaging of Alzheimer's disease (AD) for many years. Based on arguably the most dominant feature of visual examination, hippocampal volume quickly emerged as the prevailing standard MRI biomarker for quantifying AD severity from MRI (Jack et al., 1992; Kesslak et al., 1991; Scheltens et al., 1992; Seab et al., 1988). Of the two major proteinopathies characteristic of AD, ß-amyloid and tau, spatiotemporal patterns of neurodegeneration and atrophy, as well as temporal progression of clinical symptoms, correspond with the spatiotemporal pattern of tau deposition much more than with that of ß-amyloid (Jack et al., 2008; Whitwell et al., 2008). Because pathological studies suggest that the regions first affected by tau in the typical disease progression are specifically those in layers II and IV of the entorhinal cortex followed by the subiculum/CA1 of the hippocampus (Braak and Braak, 1991;

* Corresponding author at: Mayo Clinic, Diagnostic Radiology, 200 First Street SW, Rochester, MN, 55905, USA.
E-mail address: schwarz.christopher@mayo.edu (C.G. Schwarz).

[1] A portion of 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.nicl.2016.05.017

Gómez-Isla et al., 1996; Hyman et al., 1984), volume of the entorhinal cortex was also soon proposed as an AD imaging biomarker, and the relative merits of hippocampal volume versus entorhinal cortex volume have been debated (Bobinski et al., 1999; de Leon et al., 2001; Dickerson et al., 2001; Du et al., 2003; Juottonen et al., 1999; Kesslak et al., 1991; Killiany et al., 2002; Tapiola et al., 2008; Xu et al., 2000). Some studies have found superior or comparable separation or predictive power of the entorhinal cortex (Bobinski et al., 1999; Dickerson et al., 2001; Du et al., 2003; Killiany et al., 2002; Tapiola et al., 2008), while others have favored hippocampal volumes with the explanation that greater reliability of measuring hippocampal volumes compensates for its slightly later stage of being affected by the disease (Juottonen et al., 1999; Kesslak et al., 1991; Xu et al., 2000). Methods using machine learning classifiers were later proposed to analyze volume-based features of voxels/regions across the whole brain in search of a measure based on a "signature" region of interest (ROI) (i.e. a set of voxels, or a set of ROIs combined into a meta-ROI) to measure Alzheimer's disease (Fan et al., 2008; Ortiz et al., 2014; Vemuri et al., 2011, 2008; Xia et al., 2013).

When methods designed to measure in-vivo cortical thickness from MRI were introduced (Das et al., 2007; Fischl and Dale, 2000; MacDonald et al., 2000), regional thickness values, particularly in the entorhinal cortex, were quickly proposed as measures of AD severity (Bakkour et al., 2009; Dickerson et al., 2009; Fischl et al., 2009; Lerch et al., 2005). Thickness of the (whole) hippocampus is generally not considered as an option (FreeSurfer FAQ, 2015): the structure of the hippocampal cortex folds upon itself and appears bulbous, rather than thin and ribbon-like as in the rest of the cortex. Some software methods have been designed to segment hippocampal subfields, but these do not produce thickness measures (Iglesias et al., 2015; van Leemput et al., 2008). The use of cortical thickness measurements combined from multiple regions into an "AD Signature" meta-ROI has also been proposed (Dickerson and Wolk, 2012; Dickerson et al., 2009). Others have proposed longitudinal AD signature methods based on tensor-based morphometry features (Hua et al., 2009), but longitudinal measures will not be the focus of this manuscript.

Since the introduction of hippocampal volumes, these and other volume measurements have been normalized by or corrected for inter-subject variation in brain size or head size (Jack et al., 1992, 1989; Kesslak et al., 1991; Scheltens et al., 1992; Seab et al., 1988). Most commonly, head size is defined and measured as total intracranial volume (TIV), and volume measurements are applied as a ratio after dividing by TIV (Bobinski et al., 1999; Jack et al., 1992; Juottonen et al., 1999), as a residual after adjusting for TIV in a regression model (Jack et al., 2014; Voevodskaya et al., 2014), or simply as a covariate in a regression model. However, there has been no universal agreement on how to adjust volume measures for TIV, and differing methods have been shown to produce differing analyses (Hansen et al., 2015; Nordenskjöld et al., 2013; Voevodskaya et al., 2014).

Although cortical thickness and surface area seem to be biologically distinct quantities (Panizzon et al., 2009), GM volume measurements are a combination of both that is dominated by surface area much more than thickness (Winkler et al., 2010). Because surface area is highly correlated with TIV (Barnes et al., 2010), it follows that volume (which is highly correlated with surface area) is also highly correlated with TIV, while thickness (which is much less correlated with surface area) is not. This has led to previous recommendations to adjust for head size when using volume, but not thickness, measurements (Barnes et al., 2010; Westman et al., 2013). Despite these methodological differences, hippocampal volume and entorhinal cortical thickness offer similar diagnostic separability performance, and thus the degree of nuisance correlation with TIV becomes an important practical difference.

Our goal in this study is to produce a measure suitable for epidemiological assessment of disease burden over the entire age range, which includes examining the effects of sex. To assess their suitability for this purpose, we perform large-scale comparisons of several distinct techniques (volume measurements based on SPM12 and on FreeSurfer, and cortical thickness measurements based on ANTs and FreeSurfer) for creating volume- and thickness-based meta-ROI "signature" measures of AD according to a variety of desirable properties. Our evaluation criteria are the following: (a) separation of clinically normal (CN) older adults from AD patients according to clinical diagnosis; (b) (extent of) correlation with TIV in CN older adults; (c) reliability in a single-site short-time repeat-scan study of CN subjects of varying ages and in a larger, multi-site 3-month repeat-scan study of older CN subjects; and (d) correlation of these in-vivo measurements with Braak neurofibrillary tangle stage at autopsy.

2. Methods

First, we describe each of the four datasets used in this work. Additional characteristics of the subjects included in each are provided in Supplementary Table S1. All studies were approved by their respective institutional review boards and all subjects or their surrogates provided informed consent compliant with HIPAA regulations. Next, we describe all employed software pipelines. MRI scanning parameters are described later, in Section 3.

2.1. Subject characteristics

2.1.1. Mayo clinical diagnostic separability dataset

This dataset includes scans of 216 subjects from the Mayo Clinic Study of Aging (MCSA) and the Mayo Clinic Alzheimer's Disease Research Center (ADRC) studies. MCSA is an epidemiological study of cognitive aging in Rochester, Olmsted County, Minnesota (Petersen et al., 2010; Roberts et al., 2008). The ADRC study recruits and follows subjects initially seen as patients at the Mayo Clinic Behavioral Neurology practice. 108 subjects were clinically diagnosed with either AD dementia or mild cognitive impairment (MCI) according to established criteria, and 108 clinically normal control subjects were matched one-to-one to the MCI/AD subjects according to age, sex, and TIV. The MCI/AD subjects were documented to be amyloid positive and the CN subjects amyloid negative, to create an impaired group within the AD pathway and a non-impaired group not in the AD pathway (Jack et al., 2014). Amyloid positivity was determined from late uptake PET scans of each subject with Pittsburgh compound B (Klunk et al., 2004). PET image analysis was performed using a previously-described in-house automated pipeline (Senjem et al., 2005) using structural MRI to perform two-class partial volume correction and provide ROI placement (Jack et al., 2008). An amyloid-PET standardized uptake value ratio (SUVR) was calculated as the median uptake over voxels in the prefrontal, orbitofrontal, parietal, temporal, anterior cingulate, and posterior cingulate/precuneus regions divided by the median uptake from the cerebellar GM (Jack et al., 2013). A previously determined cut point of SUVR 1.4 was used to denote amyloid positive/negative (Jack et al., 2014).The clinical diagnosis criteria for normal subjects were: no cognitive complaints, normal neurological exam, no active psychiatric or neurological conditions, no psychoactive medications, and prior resolution of any previous neurological or psychiatric conditions. Subject ages were between 60 and 91 years, median 77. Two subjects, one with a diagnosis of MCI and one with AD, failed FreeSurfer processing (defined as either program failure to produce an output, or a produced output with a portion of cortical ribbon placed incorrectly) and were excluded from all FreeSurfer-based volume and thickness analyses.

2.1.2. Mayo Clinic reliability dataset

This dataset includes scans of 21 clinically normal control subjects from the Mayo Clinic MCSA/ADRC studies, pooled from two groups. It includes eight subjects, ages 30–47, with pairs of baseline and follow-up scans on the same scanner within a range of 5.9–8.2 months, and