ARTICLE IN PRESS

Psychiatry Research: Neuroimaging xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Psychiatry Research: Neuroimaging

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

Quantitative comparisons of three automated methods for estimating intracranial volume: A study of 270 longitudinal magnetic resonance images

Xiaoyan Shang[a,b], Michelle C. Carlson[c,d], Xiaoying Tang[a,b,e,f,*]

[a] Sun Yat-sen University Carnegie Mellon University (SYSU-CMU) Joint Institute of Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China
[b] Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
[c] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[d] Center on Aging and Health, Johns Hopkins University, Baltimore, MD, USA
[e] Sun Yat-sen University Carnegie Mellon University (SYSU-CMU) Shunde International Joint Research Institute, Shunde, Guangdong, China
[f] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, Guangdong, China

ARTICLE INFO

Keywords:
Total intracranial volume
Statistical Parametric Mapping
Freesurfer
Multi-atlas likelihood fusion
Magnetic resonance imaging
Automated estimation

ABSTRACT

Total intracranial volume (TIV) is often used as a measure of brain size to correct for individual variability in magnetic resonance imaging (MRI) based morphometric studies. An adjustment of TIV can greatly increase the statistical power of brain morphometry methods. As such, an accurate and precise TIV estimation is of great importance in MRI studies. In this paper, we compared three automated TIV estimation methods (multi-atlas likelihood fusion (MALF), Statistical Parametric Mapping 8 (SPM8) and FreeSurfer (FS)) using longitudinal T1-weighted MR images in a cohort of 70 older participants at elevated sociodemographic risk for Alzheimer's disease. Statistical group comparisons in terms of four different metrics were performed. Furthermore, sex, education level, and intervention status were investigated separately for their impacts on the TIV estimation performance of each method. According to our experimental results, MALF was the least susceptible to atrophy, while SPM8 and FS suffered a loss in precision. In group-wise analysis, MALF was the least sensitive method to group variation, whereas SPM8 was particularly sensitive to sex and FS was unstable with respect to education level. In terms of effectiveness, both MALF and SPM8 delivered a user-friendly performance, while FS was relatively computationally intensive.

1. Introduction

An important research topic in neuroimaging is the quantification of morphological changes of the human brain induced by normal brain development, normal aging, as well as various neuropathological and psychiatric disorders. It has been found that the variation in head size from subject to subject can be an important component of an observed inter-subject variability in various morphological measurements (such as volume, surface area, and thickness) of the whole brain and regions of interest (ROIs) (Barnes et al., 2010; Mathalon et al., 1993). To reduce variability, a standard approach is to co-vary for the effect of head size, which is usually measured as the total intracranial volume, or TIV (the volume inside the brain cranium, being the total volume of gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and meninges) (Pengas et al., 2009), when evaluating the morphological change of a specific ROI.

Studies have shown that an adjustment for TIV can considerably increase the statistical power in monitoring brain morphometry, such as the overall brain volume (Barnes et al., 2010), the GM volume (Peelle et al., 2012), and the volumes of specific ROIs (Barnes et al., 2010; Nordenskjold et al., 2013; Westman et al., 2013). Adjusting for TIV is especially important when quantifying the impact of sex on various brain morphometric measurements (Barnes et al., 2010). The scientific conclusions drawn from a morphometric study may be completely different depending on whether the effect of TIV was factored out or not. Therefore, it is a necessity to adjust the effect of TIV, for which an accurate TIV estimate is an important prerequisite.

To estimate the TIV value, the most accurate method is to manually delineate the cranial region based on 3D T1-weighted magnetic resonance (MR) images. However, this process requires trained neuroanatomists and is both exceedingly time-consuming and labor-intensive. Manual delineation also suffers from inter-operator as well as intra-

Abbreviations: ROI, Region of Interest; GM, Gray matter; WM, White matter; CSF, Cerebrospinal fluid; TIV, Total intracranial volume; MR, Magnetic resonance; SPM, Statistical Parametric Mapping; FS, FreeSurfer; MALF, Multi-atlas likelihood fusion; BHS, Brain Health Study; BECT, Baltimore Experience Corps Trial; EC, Experience Corps; MMSE, Mini-mental state examination; LDDMM, Large deformation diffeomorphic metric mapping

* Correspondence to: Guangzhou Higher Education Mega Center, No. 132, East Waihuan Road, Guangzhou 510006, China.
E-mail address: tangxiaoy@mail.sysu.edu.cn (X. Tang).

https://doi.org/10.1016/j.pscychresns.2018.02.005
Received 4 July 2017; Received in revised form 7 December 2017; Accepted 9 February 2018
0925-4927/

operator variabilities, thus inducing relatively low reproducibility (Malone et al., 2015). As such, automated and highly robust estimation methods for the TIV value are extremely useful. Two of the most popular software packages for automated TIV estimation, in the neuroimaging community, are Statistical Parametric Mapping (SPM) and FreeSurfer (FS). Various versions of both software packages have been evaluated for TIV estimation performance in previous studies. According to the results from one such study (Pengas et al., 2009), SPM outperformed FS by achieving a smaller mean absolute percentage difference between TIV values of the same participant scanned at different time points from SPM (version 5) than those produced by FS (version 3.0.2). Ridgway and colleagues (Ridgway et al., 2011) compared different versions of SPM and FS to a manual reference, including SPM (version 5), SPM (version 8), FS (version 4.5) and FS (version 5). Their results showed that SPM5 was highly variable and biased when compared to the manual measurement whereas SPM8 had largely improved upon this. The accuracy of both FS4.5 and FS5 was between that of the two SPM versions. Overall, SPM8 appeared to exhibit the best trade-off between accuracy and stability. Nordenskjöld and colleagues (Nordenskjold et al., 2013) compared SPM (version 8) and FS (version 5.1.0) with reference to manual-delineation results while also testing the influence of skull size, sex, and brain atrophy. Their results showed that both methods overestimated TIV and SPM was particularly sensitive to sex and atrophy effects whereas the accuracy of FS in TIV estimation was more likely to be affected by the skull size.

In our previous work (Tang et al., 2015a), we proposed a pipeline for skull-stripping based on our diffeomorphic multi-atlas likelihood fusion (MALF) algorithm (Tang et al., 2013), the segmentation accuracy of which has been systematically evaluated previously (Tang et al., 2015b). In this paper, we evaluated the performance of TIV estimation from MALF by taking the size of the skull-stripped brain mask as the estimated TIV and comparing its performance with that of SPM (version 8) and FS (version 4.1.0) on a large longitudinal MRI dataset obtained from a 70-participant cohort with a high risk of progressive Alzheimer's disease. Each participant had been followed up for a total of three time points (the brain of each participant being scanned every year for a continuity of three years).

One critical feature of TIV is that it remains stable over time (in adulthood) and should not be influenced by either aging or brain atrophy. In this sense, if an estimated TIV was significantly biased by atrophy, it would suggest that the specific TIV estimation method is not sufficiently robust and thus cannot be used to correct for inter-variability in terms of either head size or brain size. Therefore, in the case of a lack of the "gold standard" (the manually-measured TIV) for comparison, one can instead evaluate the change in the estimated TIV over time for the same subject. Ideally, there should be no change at all and thus this difference should be zero (Malone et al., 2015). Considering this, we applied such a criterion to our longitudinal dataset to compare MALF, SPM, and FS in terms of the automatically delivered TIV estimations. We consider a method to be "better" if that method is less susceptible to the influence of progressive brain atrophy.

In summary, this study presents results from the following investigations: (1) a comparison of the three TIV estimation methods (MALF, SPM, and FS) using four quantitative evaluation metrics; (2) statistical significance of the group differences in term of TIV estimation performance of the three methods; (3) whether or not sex, educational level, and intervention modify the TIV estimation performance of the three methods.

2. Methods

2.1. Participants and MRI parameters

Our participants were from the Brain Health Study (BHS) (Carlson et al., 2015), a trial nested within the larger Baltimore Experience Corps Trial (BECT) (Fried et al., 2013). The BECT was a sex-stratified, randomized, controlled effectiveness trial aimed at evaluating the health benefits for older adults participating in Experience Corps® (EC) Baltimore, a high-intensity volunteer service program, versus a control group offered other low-intensity volunteer opportunities. Details on sex-stratification, randomization, study design, sampling methodology, and recruitment have been described elsewhere (Fried et al., 2013). The BHS enrollment criteria have also been described in earlier studies (Carlson et al., 2015) and they included: being 60 years of age or older; having a score of 24 or higher on the mini-mental state examination (MMSE); possessing a minimum of sixth grade reading level on the Wide Range Achievement Test (Wilkinson, 1993); having right-hand dominance; being free of a pacemaker or other ferrous metals in the body; having no history of brain cancer or brain aneurism/stroke in the past year.

Among the 123 participants enrolled in BHS, only 70 underwent all three scans over the course of two years. Since our evaluation criteria are participant-specific changes across different time points, those 70 participants formed our study sample (and therefore 210 MRI scans). Excluded participants did not vary significantly from included participants in age, MMSE, sex or education (p-values greater than 0.05). Additionally, BHS participants did not differ significantly from participants in the larger BECT in terms of age, MMSE, self-reported health, and education (p-values greater than 0.05). The demographic information of the baseline sample is shown in Table 1. This study was approved by the Johns Hopkins Institutional Review Board and all participants provided written informed consent.

For each participant, all three of the serial structural MRI scans were acquired with a 3.0 T Phillips scanner (Best, the Netherlands). The imaging protocol was: high-resolution T1-weighted 3D-volume Magnetization Prepared Rapid Gradient Echo Imaging (sagittal acquisition; echo time: 3.7 ms; repetition time: 8.037 ms; flip angle: 8°; field of view = 200 mm × 256 mm × 200 mm; matrix size = 256 mm × 256 mm; and voxel size = 1 mm × 1 mm × 1 mm).

2.2. Experience corps intervention design

Participants randomized as volunteers entered one of 25 participating Baltimore City elementary schools in teams of 10–20 and were trained together, in two equal waves, a few months apart to allow schools to acclimate to the program expansion. EC volunteers received 32 h of formal training prior to placement in a designated EC school as part of a team of volunteers. Training was designed to provide necessary skills to allow volunteers to aid teachers in supporting literacy development, reading, library support, and in modeling positive communication. Volunteers were assigned by the Principal to assist a teacher with a class between kindergarten and third grade level. Those randomized to the wait-list control arm were referred to the Baltimore City Commission on Aging and Retirement Education for other low-activity volunteer opportunities in Baltimore City. Controls were wait-listed for participation in EC after two years, if interested.

Table 1
Demographic characteristics of the 70 participants from BHS that have been included in this study.

Demographic	Mean ± SD
Age (years)	67.3 ± 6.4
Sex (Male/Female)	20/50
Race (African American)	64 (91.4%)
Education (years)	14.1 ± 2.8
Education (< = High School)	25 (35.7%)
Group(Control/Intervention)	29/40
MMSE	28.4 ± 1.5

SD = standard deviation.
High school is defined as 12-year education.
MMSE = Mini Mental State Exam.