Journal of the International Neuropsychological Society (2018), **24**, 206–211.
Copyright © INS. Published by Cambridge University Press, 2017.
doi:10.1017/S1355617717000868

BRIEF COMMUNICATION

Reliability and Utility of Manual and Automated Estimates of Total Intracranial Volume

Samuel J. Crowley, Jared J. Tanner, Daniel Ramon, Nadine A. Schwab, Loren P. Hizel, AND Catherine C. Price
Clinical and Health Psychology, University of Florida, Gainesville, Florida, Gainesville, Florida

(RECEIVED December 8, 2016; FINAL REVISION July 17, 2017; ACCEPTED July 20, 2017; FIRST PUBLISHED ONLINE October 5, 2017)

Abstract

Objectives: Total intracranial volume (TICV) is an important control variable in brain–behavior research, yet its calculation has challenges. Manual TICV (Manual) is labor intensive, and automatic methods vary in reliability. To identify an accurate automatic approach we assessed the reliability of two FreeSurfer TICV metrics (eTIV and Brainmask) relative to manual TICV. We then assessed how these metrics alter associations between left entorhinal cortex (ERC) volume and story retention. **Methods:** Forty individuals with Parkinson's disease (PD) and 40 non-PD peers completed a brain MRI and memory testing. Manual metrics were compared to FreeSurfer's Brainmask (a skull strip mask with total volume of gray, white, and most cerebrospinal fluid) and eTIV (calculated using the transformation matrix into Talairach space). Volumes were compared with two-way interclass correlations and dice similarity indices. Associations between ERC volume and Wechsler Memory Scale-Third Edition Logical Memory retention were examined with and without correction using each TICV method. **Results:** Brainmask volumes were larger and eTIV volumes smaller than Manual. Both automated metrics correlated highly with Manual. All TICV metrics explained additional variance in the ERC-Memory relationship, although none were significant. Brainmask explained slightly more variance than other methods. **Conclusions:** Our findings suggest Brainmask is more reliable than eTIV for TICV correction in brain-behavioral research. (*JINS*, 2018, *24*, 206–211)

Keywords: FreeSurfer, eTIV, Brainmask, Entorhinal cortex, MRI, TICV correction

INTRODUCTION

Correcting for differences in head size improves the strength of associations between brain structures and cognitive variables of interest (Bigler & Tate, 2001). This correction is best accomplished using total intracranial volume (TICV), defined as the total volume of gray matter, white matter, meninges, and cerebrospinal fluid (CSF) within the skull. Normalizing brain structures by TICV reduces variability more effectively than normalization by cranial area or cerebral volume (Bigler et al., 2004; Whitwell, Crum, Watt, & Fox, 2001). TICV is consistent; it does not change with age or brain atrophy (Whitwell et al., 2001). For these reasons, TICV is often used to correct for neuroanatomical volumetric differences between groups (Bigler & Tate, 2001; Sargolzaei et al., 2014) and neuroanatomical–cognitive associations (Bigler et al., 2004).

Researchers can calculate TICV in several ways. The "gold standard" is manual tracing (Keihaninejad et al., 2010) or manual editing of automatically derived TICV volumes. These methods require extensive training and time, however. Automatic TICV measurement programs are a more rapid alternative, but vary in accuracy and reliability (Ridgeway, Barnes, Pepple, & Fox, 2011).

FreeSurfer, an established segmentation program, provides automatic TICV estimations. Estimated TICV (eTIV; Buckner et al., 2004) is one Freesurfer TICV measurement method. It uses the extent of the transformation needed to register a participant's image to a template to obtain an estimate TICV that includes all brain matter and CSF within the skull. This method is widely used; a search on PubMed showed over 200 citations for Buckner et al. (2004), which demonstrated the eTIV approach. There is some debate about accuracy, however. Some (e.g., Sargolzaei et al., 2014) report eTIV is more consistent with manual TICV than estimations from similar programs. Others (Nordenskjöld et al., 2015), however, report that eTIV overestimates TICV.

Correspondence and reprint requests to: Catherine Price, Clinical and Health Psychology, University of Florida, Gainesville, FL 32610. E-mail: cep23@phhp.ufl.edu

https://doi.org/10.1017/S1355617717000868 Published online by Cambridge University Press

As an alternative to eTIV, FreeSurfer can derive TICV from a native mask containing all voxels designated as gray matter, white matter, and ventricular and sulcal cerebral spinal fluid (CSF), as well as portions of subarachnoid CSF. This information is provided in a file labeled “brainmask.mgz,” referred to simply as “Brainmask” in the current study. Although not originally devised to contain all subarachnoid space, this rough skull strip can serve as an estimation of TICV. Furthermore, Brainmask does not use a template for registration, reducing potential registration-to-template errors encountered by other algorithms such as eTIV. Despite this advantage, few studies report using this method for TICV estimation, perhaps due to the program’s difficulty distinguishing between skull and CSF in T1 images during brain segmentation.

To our knowledge, no study has directly compared eTIV and Brainmask to manual TICV or examined how these two TICV measurements alter brain–behavioral associations when used as a correction variable. We designed the current study with two aims. First, we compared eTIV and Brainmask volume estimations to manual TICV. We hypothesized stronger associations between Brainmask and manual TICV relative to eTIV and manual. Second, we assessed the utility of eTIV and Brainmask metrics as TICV correction variables for expected associations between the left entorhinal cortex (ERC) and verbal memory in older adults. We hypothesized ERC volume-memory score associations would strengthen with Brainmask relative to eTIV correction.

METHODS

Participants

This was a retrospective data analysis conducted in accordance with the University of Florida’s Institutional Review Board and in compliance with the Declaration of Helsinki. Data were acquired *via* a larger federally funded investigation and included 40 individuals with Parkinson’s disease (PD) and 40 non-PD, demographically matched individuals. PD participants were included for convenience and to add behavioral range within the sample. All participants completed an MRI on a 3 Tesla scanner and a comprehensive neuropsychological assessment. Participants were required to be native English speakers, be at least 60 years of age, and have the ability to read and write. Further requirements included intact instrumental activities of daily living and non-demented *via* a comprehensive neuropsychological evaluation as part of a federally funded research investigation.

MRI Parameters and Structure Registration

Neuroimaging data were prospectively acquired with a Siemens 3T Verio scanner using an 8-channel head coil. We acquired two T1-weighted scans (176 contiguous sagittal slices, 1 mm^3 voxels, field of view (FOV) = 256 mm, sagittal orientation, 256 × 256 matrix, repetition time/echo time (TR/TE) = 2500/3.77 ms, fliop angle (FA) = 7 deg, 7/8 Partial Fourier, acquisition time 9:22) for structural brain analyses. The average of each participant’s two T1 scans was analyzed for the current study. T2-weighted images were acquired for skull segmentation (176 contiguous sagittal slices, 1 mm^3 voxels, FOV = 256, 256 × 256 matrix, sagittal orientation, TR/ TE = 3200/ 409 ms, FA = 7 deg, GeneRalized Autocalibrating Partial Parallel Acquisition, acquisition time 4:43). Images were visually examined for excessive motion, and those showing more than a moderate degree of motion were excluded from the analyses.

TICV Methods

Manual TICV

FSL Brain Extraction Tool (BET; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET), a program that estimates the inner and outer skull surface using T1 and T2 input images, provided participants’ initial TICV mask extending into the inner surface of the skull and extracted in native space. These masks were then visually inspected by two trained raters and manually edited on every sagittal slice in ITK-SNAP (www.itksnap.org) so the mask filled the space within the inner surface of the skull (i.e., filled the subarachnoid CSF space not originally captured by BET). A straight line between the superior portion of the occipital bone and the clivus served as the ventral TICV landmark. This final mask served as the semi-automated measurement of TICV, which heretofore will be referred to as “Manual.” Two expert raters established inter and intra rater reliability by creating TICV on two separate occasions for 10 brains randomly selected from the sample. Final TICV reliability metrics were excellent [Dice similarity index (DSC) intra- and inter-rater reliability >0.99; ICC intra- and inter-rater reliability >0.91]. Intra- and inter-rater reliability did not differ between PD participants and controls.

Automated TICV

All MR images were processed using FreeSurfer 5.3, producing two different TICV estimates.

eTIV

As described in Buckner et al. (2004), a 12-parameter affine transformation was used to transform each individual’s MPRAGE scan to match the Talairach template in MNI space. This transformation produced an atlas scaling factor, computed as the reciprocal of the atlas determinant. This number represents the eTIV, and is the extent of contraction or expansion required for the image to match the template. This estimation includes all CSF within the skull, including subarachnoid CSF.

Brainmask

During the initial skull strip processing in FreeSurfer, an output image called “brainmask.mgz” is created. This volumetric

https://doi.org/10.1017/S1355617717000868 Published online by Cambridge University Press

mask, traditionally used as a skull strip for further analysis, provides a liberal outline of the brain with some subarachnoid CSF included, and contains all voxels in the brain determined to be grey matter, white matter or CSF. The total voxel count is used as a metric of TICV. Unlike eTIV, Brainmask creates a three-dimensional (3D) volume file that can be used for spatial overlap analyses with manual TICV. For this analysis, we did not manually edit Brainmask.

ERC Acquisition

Structural T1 scans were processed *via* FreeSurfer 5.3 (Segonne et al., 2004; Fischl et al., 2002; http://surfer.nmr.mgh.harvard.edu/). An averaged brain with enhanced gray-white contrast and increased signal-to-noise was aligned to the MNI152 template brain using a linear registration technique with 6 degrees of freedom (FLIRT; Jenkinson, Bannister, Brady, & Smith, 2002) to correct for head tilt and align participants' brains along the anterior commissure–posterior commissure axis. Freesurfer's automatic ERC volumes were not used due to their limited reliability (McCarthy et al., 2015). Instead, an expert reliable rater (J.T.; intra-rater DSC > 0.8; inter-rater reliability DSC > 0.80) manually traced the entorhinal cortices using published criteria (Insausti et al., 1998). Starting 2 mm posterior to the appearance of the temporal stem, the lateral wall of the parahippocampal gyrus was traced between the sulcus semiannularis and the collateral sulcus, descending into the collateral sulcus at varying depths depending on its overall depth. This method produces volumes that have been associated with verbal memory performance in older adults with and without Parkinson's disease (Price et al., 2010; Tanner et al., 2015).

Verbal Memory Measure

The Logical Memory (LM) subtest of the Wechsler Memory Scale-Third Edition (WMS-III; Wechsler, 1997) assessed verbal learning and memory. The primary outcome variable of interest was number of items recalled at delayed recall divided by the number of items recalled at immediate recall, known as Logical Memory Retention (LMret). The values used in the analysis were *Z*-scores derived from the WMS-III normative sample, which is stratified by age group.

Statistical Analysis

Reliability

Lin's concordance correlation coefficient (CCC; Lawrence & Lin, 1989) assessed reliability for eTIV and Brainmask relative to Manual. A DSC examined overlap between Brainmask and Manual. DSC could not be calculated between eTIV and Manual because eTIV is not a 3D volume. A Fisher *Z* statistic examined the extent of differences between correlation coefficients.

A Bland-Altman analysis was also used to assess the degree of disagreement between the automatic and manual TICV measures (Altman & Bland, 1983). Bland-Altman plots show differences between two measures by plotting the average of the two measures *versus* the difference between the two measures. Lines are then added for the mean difference and at ±1.96 standard deviations from the mean difference, termed the "agreement range." This method shows the range of values in which the majority of error is likely to occur. It also shows whether measurement errors are consistent at all values (i.e., at all values of TICV) or if it is uneven with different values (i.e., shows more inaccuracy at lower or higher values of TICV).

ERC-LMret associations

Three-step hierarchical regressions examined TICV metrics contributed to ERC and LMret. Step one included PD status as a nuisance variable. Step two included ERC raw volume predicting retention. Step three examined how each TICV metric altered the overall model strength and ERC-LMret association.

RESULTS

Participants

The two participant groups did not statistically differ in age, education, gender, or general cognition as measured by the Mini Mental State Examination (Folstein, Folstein, & McHugh, 1977) (Supplementary Table 1). PD participants had a significantly larger mean TICV than controls according to all three TICV metrics.

eTIV/Brainmask Volumes and Reliability to Manual TICV

Results of the reliability analyses are shown in Table 1. Relative to manual TICV, Brainmask volumes were significantly larger (1.88% ± 3.56% difference; t(79) = 4.67; $p < .01$), while eTIV volumes were significantly smaller than Manual (–6.91% ± 11.04% difference; t(79) = 5.67; $p < .01$) and showed a greater variability than Manual or Brainmask estimations. On an individual level, the largest percent difference between an eTIV and Manual estimation for a single image was 40%, while the largest difference between a Brainmask and Manual estimation for a single image was 12%.

Brainmask volume estimations strongly correlated with Manual (r[80] = 0.94; CCC = 0.92; $p < .01$). There was also high spatial and volumetric overlap between Manual and Brainmask (DSC = 0.95). Correlations between eTIV and Manual were weaker but significant (r[80] = 0.82; CCC = 0.68; $p < .01$). Brainmask/manual and eTIV/manual coefficient values were statistically different from one another (Fisher $Z = 3.61$; $p < .01$).

https://doi.org/10.1017/S1355617717000868 Published online by Cambridge University Press

Table 1. Reliability statistics comparing Brainmask and eTIV to manual TICV

	Manual	Brainmask	eTIV
% Diff from manual	NA	1.88 ± 3.56 (–0.08–0.12)	– 6.91 ± 11.04 (–0.40–0.08)
CCC	NA	0.92	0.68
DSC	NA	0.95	NA
Volume	1665 (166)	1696 (16)	1568 (256)
ERC proportion	.08 (0.01)	.073 (0.01)	.08 (0.02)
ß value	0.27*(Δß = 0.04)	0.29* (Δß = 0.06)	0.25* (Δß = 0.03)

Note. %Diff = percent difference between FreeSurfer measures; CCC = concordance correlation; DSC = Dice similarity coefficient; NA = not applicable. The range of % difference scores is provided in parentheses. $p < 0.001$ for all correlations. Volume is presented in cubic centimeters, and ERC is presented as percentage of TICV. SD is presented in parentheses. Beta weights when controlling for TICV using all three methods, ß = beta weights controlling for TICV with each method. Δß = change in ß from no TICV correction. *$p <$ 0.05.

The results of the Bland-Altman plots (Figure 1) showed Brainmask has a much smaller 95% range of agreement (Figure 1a; –5.09% to 8.85% difference) than eTIV (Figure 1b; –28.55% to 14.73% difference), indicating QJ;eTIV has a higher degree of error. The most severe underestimations by eTIV occurred in individuals with smaller TICVs, indicating a greater chance of inaccuracy of eTIV when measuring smaller TICVs. Correspondingly, there was a strong association between TICV and eTIV estimation error (r = – 0.66; $p < .001$). In contrast, the correlation between TICV and Brainmask error was not significant (r = – 0.17).

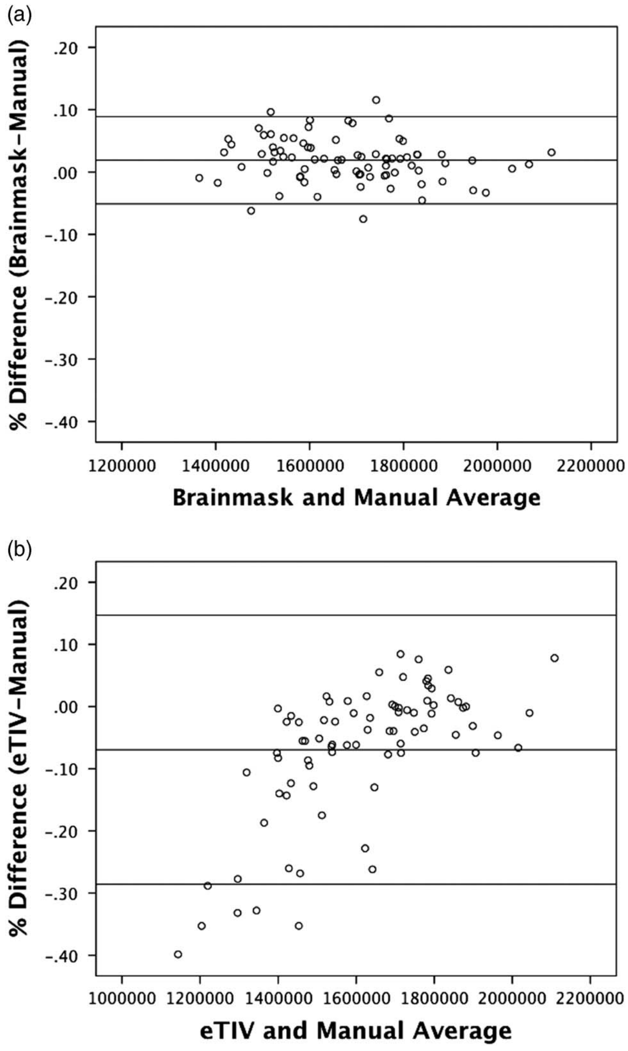

Fig. 1. Results of the Bland-Altman plots: X-axis represents average TICV in cubic mm. Y-axis represents % difference between automatic TICV methods and manual TICV.

ERC-LM Retention Associations

Results of this analysis are displayed in Table 1. Step 1 of the hierarchical regression analysis showed a significant effect of group (R^2 = 0.07; p = .02). Step 2 of the hierarchical model found ERC raw volume positively associated with LMret (ΔR^2 = 0.23; p = .04). Adding in manual TICV to the model increased the ERC-LMret beta weights (ß = 0.27; p = .02), despite no change to the overall model (ΔR^2 = 0.02; p = .20). The same pattern was found when using Brainmask (ΔR^2 = 0.04; p = .06; ß = 0.29; p = .01) as well as eTIV (ΔR^2 = 0.02; p = .26; ß =0.25; p = .02).

DISCUSSION

Two Freesurfer TICV metrics differed significantly in size and association to the traditional gold standard hand drawn approach. Each TICV metric explained additional variance for ERC-memory associations with Brainmask explaining the most variance. These findings show TICV metrics are not all alike and choice of metric deserves consideration, particularly for brain–behavioral investigations.

In our sample, Brainmask volumes were larger than manual TICV, while eTIV volumes were smaller. The overestimation in Brainmask volumes may reflect FreeSurfer's difficulty discriminating brain from skull, resulting in small regions of skull being included in the final TICV measurement (Segonne et al., 2004). Although larger, Brainmask volumes are prefereable to eTIV volumes for at least two reasons. First, Brainmask showed a statistically stronger association with Manual. Second, Bland-Altman analyses revealed larger errors in eTIV; there were greater estimation errors in individuals with smaller manual TICVs compared to those with larger manual TICVs. We also

https://doi.org/10.1017/S1355617717000868 Published online by Cambridge University Press

note that adusting for TICV using Brainmask rather than eTIV resulted in larger positive ERC-LMret beta weights.

We recognize study limitations. The significance of TICV correction was likely minimized by our study sample of non-demented older adults with little to no brain atrophy; we may have observed a more striking difference in TICV metric if participants with brain volume loss (e.g., dementia) had been included in this investigation. In addition, given that Brainmask does not theoretically include portions of subarachnoid space, it may be a less accurate measure of TICV in older populations when atrophy results in an increase in subarachnoid space. We did not run the FreeSurfer recon-all script also importing T2-weighted images, which might have improved Brainmask estimations. Finally, Brainmask is theoretically more related to total brain volume than TICV, despite our findings showing it is a good estimation of TICV. For all of these reasons, future studies should assess the association between Brainmask, other TICV metrics, and total brain volume in neurodegenerative disorders.

Given these findings, we provide the following recommendations to non-dementia researchers. TICV estimatations with eTIV may need manual editing, particularly for poor registrations to templates. If researchers use eTIV, we encourage reporting steps taken to correct registration errors (for example, a description of quality control checks and a count of images requiring manual manipulation). Given the time demands of manual editing, researchers may wish to use Brainmask as the metric appears accurate to manual measurements and requires little or no manual intervention with high quality scans. A cons to working with Brainmask is that potential underrepresentation of subarachnoid CSF space. As newer software updates become available to the research community, it is imperative not only to assess their reliability relative to manual volumes, but also their effectiveness in controlling for head size when assessing brain volume and cognition associations.

ACKNOWLEDGMENTS

Samuel J. Crowley, Jared J. Tanner, Daniel Ramon, Nadine A. Schwab, Loren P. Hizel, and Catherine C. Price declare that they have no conflicts of interest. We are most grateful to the time and efforts of the research participants. We are also grateful to Tony Mancuso, Ph. D., Chair, Department of Radiology, University of Florida, Gainesville, Florida, for help acquiring use of the Siemens 3T Verio Scanner. This work was supported by the NIH grants K23NS60660, NS082386NR014810, as well as the National Institutes of Health Clinical and Translational Science Award program, grants UL1TR001427, KL2TR001429 and TL1TR001428. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary material

Supplementary material can be found online. Please visit https://doi.org/10.1017/S1355617717000868.

REFERENCES

Altman, D.G., & Bland, J.M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, 307–317.

Bigler, E.D., & Tate, D.F. (2001). Brain volume, intracranial volume, and dementia. *Investigations in Radiology*, *36*, 539–546.

Bigler, E.D., Neeley, E.S., Miller, M.J., Tate, D.F., Rice, S.A., Cleavinger, H., … Welsh-Bohmer, K. (2004). Cerebral volume loss, cognitive deficit and neuropsychological performance: Comparative measures of brain atrophy: I. Dementia. *Journal of the International Neuropsychological Society*, *10*(03), 442–452.

Buckner, R.L., Head, D., Parker, J., Fotenos, A.F., Marcus, D., Morris, J.C., & Snyder, A.Z. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: Reliability and validation against manual measurement of total intracranial volume. *NeuroImage*, *23*, 724–738.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Montillo, A. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Insausti, R., Juottonen, K., Soininen, H., Insausti, A.M., Partanen, K., Vainio, P., … Pitkänen, A. (1998). MR volumetric analysis of the human entorhinal, perirhinal, and temporopolar cortices. *AJNR American Journal of Neuroradiology*, *19*(4), 659–671.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841.

Keihaninejad, S., Heckemann, R.A., Fagiolo, G., Symms, M.R., Hajnal, J.V., & Hammers, A. (2010). A robust method to estimate the intracranial volume across MRI field strengths (1.5T and 3T). *NeuroImage*, *50*, 1427–1437.

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268.

Malmberg, F., Larsson, E.M., Simmons, A., Ahlström, H., Johansson, L., & Kullberg, J. (2015). Intracranial volume normalization methods: considerations when investigating gender differences in regional brain volume. *Psychiatry Research: Neuroimaging*, *231*(3), 227–235.

McCarthy, C.S., Ramprashad, A., Thompson, C., Botti, J.A., Coman, I.L., & Kates, W.R. (2015). A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in Neuroscience*, *9*, 379.

Nordenskjöld, R., Malmberg, F., Larsson, E.M., Simmons, A., Ahlström, H., Johansson, L., & Kullberg, J. (2015). Intracranial volume normalization methods: considerations when investigating gender differences in regional brain volume. *Psychiatry Research: Neuroimaging*, *231*(3), 227–235.

Price, C.C., Wood, M.F., Leonard, C.M., Towler, S., Ward, J., Montijo, H., … Schmalfuss, I. (2010). Entorhinal cortex volume in older adults: Reliability and validity considerations for three published measurement protocols. *Journal of the International Neuropsychological Society*, *16*(05), 846–855.

Ridgway, G., Barnes, J., Pepple, T., & Fox, N. (2011). Estimation of total intracranial volume; a comparison of methods. *Alzheimers Dementia*, *7*(4), S62–S63.

https://doi.org/10.1017/S1355617717000868 Published online by Cambridge University Press

Sargolzaei, S., Goryawala, M., Cabrerizo, M., Chen, G., Jayakar, P., Duara, R., … Adjouadi, M. (2014). Comparative reliability analysis of publicly available software packages for automatic intracranial volume estimation. *Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE: 2014 IEEE*. 2342–2345.

Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., & Fischl, B. (2004). A hybrid approach to the skull-strip ping problem in MRI. *NeuroImage*, *22*(3), 1060–1075.

Tanner, J.J., Mareci, T.H., Okun, M.S., Bowers, D., Libon, D.J., & Price, C.C. (2015). Temporal lobe and frontal-subcortical dissociations in non-demented Parkinson's disease with verbal memory impairment. *PLoS One*, *10*, e0133792.

Wechsler, D. (1997). *Wechsler memory scale: Third edition*. San Antonio, TX: The Psychological Corporation.

Whitwell, J.L, Crum, W.R., Watt, H.C., & Fox, N.C. (2001). Normalization of cerebral volumes by use of intracranial volume: –cations for longitudinal quantitative MR imaging. *AJNR American Journal of Neuroradiology*, *22*, 1483–1489.

https://doi.org/10.1017/S1355617717000868 Published online by Cambridge University Press