Validity Assessment of an Automated Brain Morphometry Tool for Patients with De Novo Memory Symptoms

F. Rahmani, S. Jindal, C.A. Raji, W. Wang, A. Nazeri, G.G. Perez-Carrillo, M.M. Miller-Thomas, P. Graner, B. Marechal, A. Shah, M. Zimmermann, C.D. Chen, S. Keefe, P. LaMontagne, and T.L.S. Benzinger

ABSTRACT

BACKGROUND AND PURPOSE: Automated volumetric analysis of structural MR imaging allows quantitative assessment of brain atrophy in neurodegenerative disorders. We compared the brain segmentation performance of the AI-Rad Companion brain MR imaging software against an in-house FreeSurfer 7.1.1/Individual Longitudinal Participant pipeline.

MATERIALS AND METHODS: T1-weighted images of 45 participants with de novo memory symptoms were selected from the OASIS-4 database and analyzed through the AI-Rad Companion brain MR imaging tool and the FreeSurfer 7.1.1/Individual Longitudinal Participant pipeline. Correlation, agreement, and consistency between the 2 tools were compared among the absolute, normalized, and standardized volumes. Final reports generated by each tool were used to compare the rates of detection of abnormality and the compatibility of radiologic impressions made using each tool, compared with the clinical diagnoses.

RESULTS: We observed strong correlation, moderate consistency, and poor agreement between absolute volumes of the main cortical lobes and subcortical structures measured by the AI-Rad Companion brain MR imaging tool compared with FreeSurfer. The strength of the correlations increased after normalizing the measurements to the total intracranial volume. Standardized measurements differed significantly between the 2 tools, likely owing to differences in the normative data sets used to calibrate each tool. When considering the FreeSurfer 7.1.1/Individual Longitudinal Participant pipeline as a reference standard, the AI-Rad Companion brain MR imaging tool had a specificity of 90.6%–100% and a sensitivity of 64.3%–100% in detecting volumetric abnormalities. There was no difference between the rate of compatibility of radiologic and clinical impressions when using the 2 tools.

CONCLUSIONS: The AI-Rad Companion brain MR imaging tool reliably detects atrophy in cortical and subcortical regions implicated in the differential diagnosis of dementia.

ABBREVIATIONS: AD = Alzheimer disease; AIRC = AI-Rad Companion; CDR = Clinical Dementia Rating; DLB = dementia with Lewy bodies; FTD = frontotemporal dementia; FS = FreeSurfer; GDS = Geriatric Depression Scale; ICC = intraclass correlation coefficient; ILP = Individual Longitudinal Participant; OASIS = Open Access Series of Imaging Studies; TIV = total intracranial volume

Standard of care for any cognitive or memory issues includes structural MR imaging of the brain.[1] Beyond its utility to exclude anatomic or pathologic abnormalities, structural brain MR imaging enables volumetric quantification of different brain structures that are affected by neurodegenerative diseases that cause cognitive impairment. FreeSurfer (FS; https://surfer.nmr.mgh.harvard.edu/) is the most commonly used volumetric analysis tool, using an automated ROI-based algorithm to generate thickness, surface areas, and volumes for 68 different cortical and subcortical regional volumes.[2-4]

Due to the detailed scale of the FS output, it is often incorporated into further processing to summarize the results into meaningful metrics for different diagnostic purposes, namely dementia. One such pipeline is a 2-step processing pipeline consisting of FS Version 7.1.1 processing of structural T1 images followed by the Individual Longitudinal Participant (ILP) software Version 2.0 (herein and after referred to as the FS/ILP pipeline;[5] for volumetric brain assessment. The time-exhaustive nature of this research-standard pipeline, which includes generation of the FS output (between 6 and 12 hours),

Received June 30, 2022; accepted after revision January 9, 2023.

From the Mallinckrodt Institute of Radiology, Division of Neuroradiology (F.R., S.J., C.A.R., W.W., A.N., G.G.P.-C., M.M.M.-T., C.D.C., S.K., P.L., T.L.S.B.) and Charles F. and Joanne Knight Alzheimer Disease Research Center (F.R., S.J., C.A.R., W.W., A.N., C.D.C., T.L.S.B.), Washington University in St. Louis, St. Lous, Missouri; Siemens Medical Solutions (P.G., B.M., M.Z.), Malvern, Pennsylvania; Advanced Clinical Imaging Technology (P.G., B.M., M.Z.), Siemens Healthcare, Lausanne, Switzerland; Department of Radiology (P.G., B.M., M.Z.), Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland; LTS5, École Polytechnique Fédérale de Lausanne (P.G., B.M., A.S., M.Z.), Lausanne, Switzerland; and Siemens Healthcare (P.G., B.M., M.Z.), Erlangen, Germany.

Please address correspondence to Tammie L.S. Benzinger, MD, PhD, Mallinckrodt Institute of Radiology, Washington University School of Medicine in St. Louis, 510 South Kingshighway Boulevard, Campus Box 8131, St. Louis, MO 63110; e-mail: benzingert@wustl.edu; @BenzingerNeuro1; @tlsmit1

Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A7790

visual inspection, and potential manual editing and recalculation steps of the output, limits its applicability in high-throughput clinical settings. Siemens has developed the FDA-cleared AI-Rad Companion (AIRC; Siemens) brain MR software that enables volumetric quantification of main cortical and subcortical structures in a scale of a few minutes (herein after referred to as the AIRC tool). We therefore aimed to investigate the validity of the AIRC tool in a clinical context through the following: 1) assessment of the correlation, consistency, and agreement of volumetric measurements generated by the AIRC tool versus those produced by the FS/ILP pipeline; 2) assessment of the sensitivity of the AIRC output in the detection of volumetric abnormalities associated with neurodegenerative causes of dementia, compared with the output from the FS/ILP as a reference standard; and 3) assessment of the potential effect of any discrepant finding between the 2 tools on the final impression made by a radiologist.

MATERIALS AND METHODS

Participants

Participants were randomly selected from the Open Access Series of Imaging Studies 4 (OASIS-4) cohort, which is publicly accessible through the OASIS brain website: https://central.xnat.org/. Participants were included under the following circumstances: 1) They were referred for clinical assessment due to a de novo cognitive symptom, 2) were 45 years of age or older, and 3) had a structural T1-weighted MR imaging study within a maximum of 1 year of the initial assessment. Diagnosis of dementia was made on the basis of the clinical assessment and a battery of cognitive tests, including the Clinical Dementia Rating (CDR)[6] and Mini-Mental State Examination.[7] If dementia was present, an etiologic diagnosis was further determined on the basis of clinical practices for Alzheimer disease (AD), posterior cortical atrophy, dementia with Lewy bodies (DLB), frontotemporal dementia (FTD), and vascular cognitive impairment.[8-11] This diagnosis was made by a neurologist clinician at the end of the recruitment visit and before any imaging assessment. Reflecting the proportion of each disease category in the OASIS-4 cohort, the current sample comprised a random selection of 15 individuals with AD, including 5 participants with early-onset AD; 10 participants with non-neurodegenerative conditions; such as subjective cognitive impairment in the absence of clinical dementia, mood disorders, polypharmacy, and sleep disorders; 5 with posterior cortical atrophy; 5 with DLB; 5 with FTD, and 5 with vascular cognitive impairment (Online Supplemental Data). The 15-item version of the Geriatric Depression Scale (GDS) was used to screen participants for the presence of depressive symptoms in which a cutoff score of 5 has shown 92% sensitivity (Online Supplemental Data).[12]

This study was conducted using a research agreement between Washington University School of Medicine in St. Louis and Siemens Medical Solutions USA and was reviewed and approved by the institutional review board of Washington University in Saint Louis School of Medicine (IRB No. 201912172).

Image Data Collection

The 3D T1WI MPRAGE and T2-weighted FLAIR images were acquired on a 3T scanner (Magnetom Skyra; Siemens) using a TR/TE = 2300/2.95 ms, TI = 900 ms, flip angle = 9°, section thickness = 1 mm, and FOV = 256 × 256 for the T1WI scans, and a TR/TE = 900/81 ms, TI = 2500 ms, flip angle = 150°, section thickness = 5 mm, within a 256 × 256 FOV for FLAIR scans. SWI was performed in the same session and used the following parameters: TR/TE = 27/20 ms, flip angle = 15°, section thickness = 2.4 mm, FOV = 256 × 256.

In all subsequent analyses, the absolute volumes refer to raw estimates produced by each tool, normalized volumes refer to absolute volumes divided by their corresponding estimated total intracranial volume (TIV), and standardized volumes refer to *z* scores calculated by comparing the normalized volumes with their respective normative database.

AIRC Tool

The AIRC Brain MR tool creates brain morphometry reports using a T1WI MPRAGE series and through a tissue-wise segmentation model, resulting in a considerably reduced computation time (2–5 minutes) compared with other segmentation software such as FreeSurfer.[3,13] This tool produces volumes of 25 different brain regions in both hemispheres (50 total) and compares them with age- and sex-matched normative data from a healthy population. Results are presented as a labeling report consisting of a label map showing the segmentation results (Fig 1*A*); a deviation report consisting of a deviation map and the corresponding standardized volumes for each region; and a list of evaluated volumes and their corresponding TIV-normalized measures displayed alongside the 10th–90th percentile normative ranges based on the participant's age and sex group. Regions with normative volumes that are outside this range are indicated by an asterisk (Fig 1*B*, -*C*).[13] Once processed, a visual quality check of the labeling and deviation results is performed to ensure consistent delineation of different cortical and subcortical regions. All of the 45 scans passed this quality control.

Normative Range Analyses. The normative database for the AIRC tool consists of T1-weighted MR images of 303 healthy subjects, including 50.8% men (median age, 73.25 years; age range, 19–91 years). Scans were collected from 2 cohorts: 1) the Alzheimer Disease Neuroimaging Initiative (ADNI; https://adni.loni.usc.edu/) using standard protocols for participant selection and scanning protocol,[14,15] and 2) Siemens collection of the MR imaging scans following the ADNI selection guidelines.

Normative ranges were calibrated on the respective healthy absolute volumes estimated by the AIRC using a log-linear regression model, taking into account the confounding effects of age and sex as covariates.[13] The deviation map offers a color-coded preview of the amount of deviation based on *z* score estimates of each structure.[13]

The FS/ILP Pipeline

FS Segmentation. T1-weighted images were processed with FS Version 7.1.1 and resampled to 1 × 1 × 1 mm resolution for volumetric segmentation and cortical reconstruction.[3] Regional volumes and cortical thicknesses were derived for 68 cortical and 40 subcortical regions in the left and right hemispheres after quality control of FS output through visual inspection.

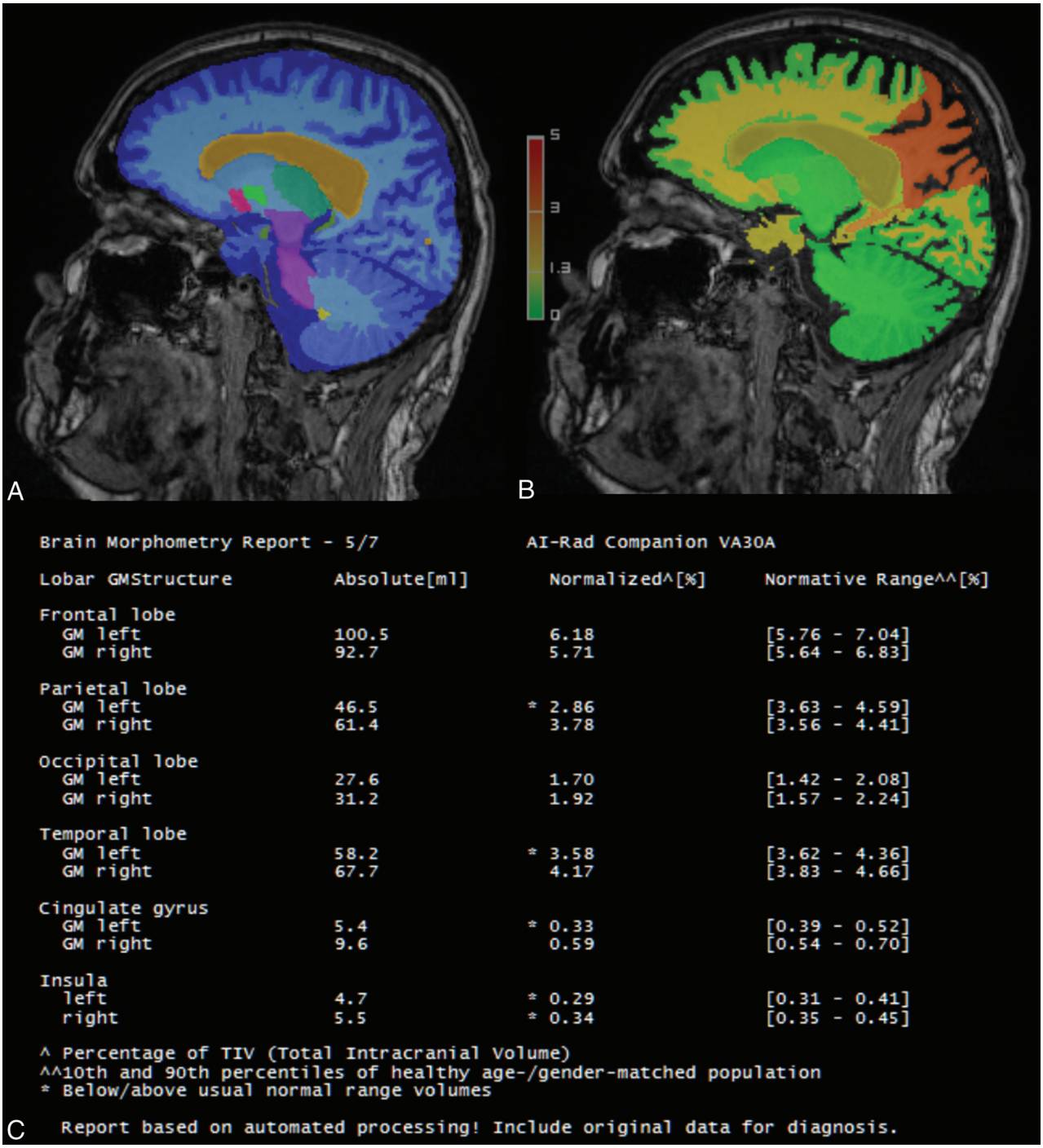

FIG 1. AIRC brain MR imaging tool volumetric output for a 60-year-old male participant with early-onset Alzheimer disease. Labeling map (*A*), deviation map (*B*), and 1 page of the numeric report (*C*). *Asterisk* indicate values outside the normative 10th–90th percentile range for participants age and sex.

Normative Range Analyses and ILP Report Generation. Once generated, FS volumes were compared with the ILP normative data sets consisting of T1-weighted MR imaging scans of 383 cognitively healthy participants assembled from 2 different sources: 1) 249 participants 38 to 88 years of age from the recently released publicly available data in OASIS-3,[16] and 2) 134 mutation-negative participants 18 to 58 years of age from the control group of the Dominantly Inherited Alzheimer Network data set (https://dian.wustl.edu/; previously published as a Normal Aging Cohort by Koenig et al[17]).

The ILP pipeline calculates a number of summary metrics based on TIV-normalized volumes and cortical thicknesses from FS output: frontal lobe cortical thickness, parietal lobe cortical thickness, occipital lobe cortical thickness, left and right hippocampal volume, left and right FTD cortical thickness (a summary measurement of cortical regions affected by frontotemporal dementia), total lateral ventricular volume, and the ratio of lateral ventricular volume to cerebral volume (Online Supplemental Data). These summary metrics are then used to generate a regression model that demonstrates

age-adjusted ranges for these volumes and thicknesses using the ILP normative data sets, forming the ILP Report.[5] With each T1WI scan processed through the FS/ILP pipeline, the above summary metrics are calculated and plotted on their corresponding ILP graph, in which the x-axis represents participants' ages and the y-axis shows the respective thickness or volume summary metric.

Analytical Approach and Statistics

Statistical analyses were performed by using R software Version 4.0.5 (http://www.r-project.org/). The purpose of these analyses was the following: 1) to assess the magnitude of the correlation, consistency, or agreement between measurements from each tool, and 2) to evaluate the sensitivity and specificity of the AIRC tool compared with FS/ILP as a reference standard. The Pearson correlation and intraclass correlation statistics were used to compare the absolute and normalized regional volumes and *z* scores derived from the FS/ILP and AIRC tools and their respective normative data sets. When necessary, a summation of various FS-based cortical segmentation volumes was calculated to match the lobar cortical volumes reported by the AIRC tool as detailed in the Online Supplemental Data.[18] The Pearson correlation coefficient and intraclass correlation coefficients (ICCs) in agreement and consistency and their respective *P* values were calculated by using the "corr" and "icc" functions, respectively. Pearson correlation coefficient values of <0.3, between 0.3 and 0.5, and >0.5 were considered to indicate small, moderate, and large correlations, while ICC values <0.5, between 0.5 and 0.7, between 0.7 and 0.9, and >0.9 were considered to indicate poor, moderate, good, and excellent agreement or consistency.[19,20] Additional details on the definition of these terms can be found in the Online Supplemental Data. Normal distribution of the variables was tested using the Kolmogorov-Smirnov goodness-of-fit test. *P* values $<$ .05 after correction for multiple comparisons using the Benjamini-Hochberg false-discovery-rate correction rejected the null assumption.[21] We further performed paired statistics to extract mean differences and the resulting effect sizes between volumes measured by each tool, as detailed in the Online Supplemental Data.

We compared the rates of detection of abnormal findings through comparison of final reports generated by each tool and by using the "chisq.test" function in R. The T1WI MPRAGE scans were evaluated by 3 board-certified neuroradiologists (W.W., C.A.R., and A.N.) with or without additional volumetric information provided by the AIRC or FS/ILP tools. Each participant was rated 3 times with 3 different methods, once using only the T1-weighted image (MPRAGE_Only), once after adding the FS/ILP output (MPRAGE+ILP), and once after adding the AIRC output (MPRAGE+AIRC). Raters independently assessed all 45 cases so that each participant was randomly evaluated by using one of the above 3 methods by each rater. The raters were asked to indicate radiologic impressions in a stepwise manner indicating the following: 1) whether there were any structural abnormalities related to the patient's cognitive symptoms, 2) whether the observed abnormalities were symmetric and lobar, and 3) whether the abnormalities pointed to a specific neurodegenerative entity (AD, posterior cortical atrophy, DLB, FTD, vascular cognitive impairment). The rate of compatibility between radiologic impressions and clinical diagnoses was calculated as percentages for each method as detailed in the Online Supplemental Data and compared across methods using the "aov" and "TukeyHSD" functions in R.

RESULTS

The Online Supplemental Data demonstrate a summary of clinical and demographic features of the study population including their cognitive status assessed through CDR, CDR sum of boxes and the Mini-Mental State Examination scores, and the presence or absence of depressive symptoms based on the 15-item GDS score. The GDS score ranged between 0 and 6 across participants in all diagnosis groups, while participants with non-neurodegenerative causes for their cognitive symptoms were more likely to have a GDS score of $\geq$5, compared with participants diagnosed with neurodegenerative conditions. Figure 2 demonstrates the results of comparisons between the 2 tools based on the absolute, normalized, and standardized regional volumes.

Absolute Volumes

There was a large, positive relationship (Pearson correlation coefficient) and excellent-to-good consistency (ICC-consistency) between measured absolute volumes of the brain, cerebellum, lateral ventricles, and putamen and between the AIRC tool and FS. Absolute volumes of the frontal, parietal, occipital, and temporal lobes and the hippocampal volumes demonstrated moderate-to-poor agreement (ICC-agreement) and consistency between by the AIRC tool and the FS/ILP pipeline. Thalamic absolute volumes demonstrated the weakest consistency between the 2 tools and no significant agreement (ICC-a) or correlation (ICC-c) (Fig 2).

Normalized Volumes

When we compared volumes normalized to the TIV, a large, positive correlation and moderate consistency were observed in both cortical and subcortical regional volumes between the 2 tools, while there was an increase in the correlation coefficients for most regions (Fig 2). Normalized brain, cerebellum, and lateral ventricular volumes demonstrated an excellent consistency and agreement when compared between the 2 tools. There was no significant agreement (ICC-a) in the normalized volumetric measurements of the bilateral frontal lobes, thalami, and putaminal regions (Fig 2).

Standardized Volumes

Once volumes were transformed to standardized *z* scores, correlation and consistency were moderate among *z* scores of the 4 main cortical lobes as well as the bilateral hippocampi (Fig 2). There was no significant agreement (ICC-a) in the regional *z* scores except in the bilateral pallidum, putamen, insula, and lateral ventricles (Fig 2).

Comparing the Diagnostic Utility of Outputs from the FS/ILP versus the AIRC Tools

We compared the performance of the AIRC tool and FS/ILP pipelines through comparison of the final report generated by the 2 tools. Cutoff points indicating abnormal regional values were either above +2 SDs ($>$97.5th percentile, for ventricular volumes and the ventricle/cerebrum ratio) and below −2 SDs ($<$2.5th percentile, for all other region/metrics) in the FS/ILP

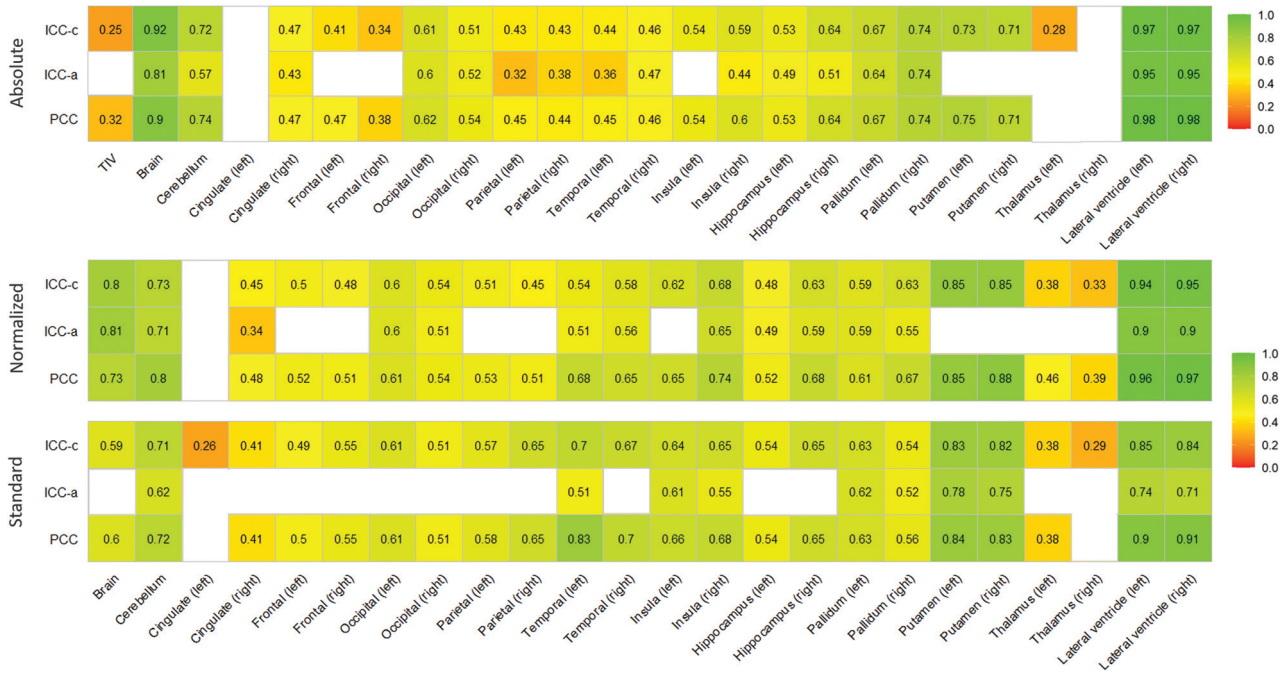

FIG 2. Comparing the Pearson and intraclass correlation between volumetric measurements produced by the AIRC-versus-FS/ILP tools. Panels demonstrate correlation coefficients for raw volumes (*top*), volumes normalized to TIV (*middle*), and standardized (*z* score) volumes (*bottom*). Blank cells demonstrate absence of statistically significant correlation between the 2 tools. ICC-c indicates Intraclass correlation coefficient-consistency; ICC-a, ICC-agreement; PCC, Pearson's correlation coefficient.

output, corresponding to >90th percentile (for ventricular volumes and the ventricle/cerebrum ratio) and <10th percentile (for all other region/metrics) in the AIRC tool output (Online Supplemental Data).

The Online Supplemental Data show a comparison between rates of detection of abnormal findings by the 2 tools, considering the FS/ILP pipeline as a reference standard. Note that in this step and for the main lobes (frontal, parietal, temporal, and occipital), volume-based *z* scores from the AIRC were compared with thickness-based *z* scores from the FS/ILP tool. This step was unlike the previous steps in which volumes generated by each tool were compared with each other. The AIRC tool had a high specificity in the detection of volumetric abnormalities, ranging from 90.6% in detecting enlarged lateral ventricles to 100% in detecting concurrent frontal and temporal atrophy (FTD thickness in FS/ILP output). Sensitivity ranged between 64.3% and 100%, with the lowest rate detected in the comparison between concurrent frontal and temporal lobe atrophy in the AIRC output and FTD thickness in the FS/ILP output. AIRC was 94.4% specific and 78% sensitive in the detection of hippocampal atrophy compared with the FS/ILP pipeline.

Equal Rate of Compatible Diagnoses Using the FS/ILP versus AIRC Tools

Each participant was independently evaluated 3 times, each time based on one of the following combinations of methods: MPRAGE_only, MPRAGE+FS/ILP, and MPRAGE+AIRC. Impressions made by the neuroradiologists were then compared with the diagnoses made by the clinician as the reference standard and marked as either compatible or incompatible (Online Supplemental Data).

Our findings indicated no difference in the rate of compatibility with clinical impressions among radiologic impressions made on the basis of the MPRAGE+ILP or MPRAGE+AIRC methods (χ^2 *P* value > .05). Even among participants with a known neurodegenerative diagnosis (35 of 45), there were no significant differences in the rate of detection of abnormality, symmetric and lobar atrophy, or the presence/absence of a neurodegenerative cause between 2 methods (Online Supplemental Data). Finally, we could not detect any difference in the rate of compatibility of the clinical diagnoses with the radiologic impressions made on the basis of either of the tools compared with the impressions made in the absence of quantitative volumetric assessment (based on the T1-weighted structural image [MPRAGE_only]).

DISCUSSION

We compared the AIRC brain MR imaging tool, a commercially available volumetric brain assessment software, with the standard publicly accessible FS/ILP pipeline. We used a sample of 45 individuals with a de novo memory symptom to investigate the effect of any potential discrepancy between the 2 tools. We found the following: 1) volumetric measurements produced by the FS/ILP and AIRC tools were largely correlated and moderately consistent in most cortical and subcortical structures, a relationship that improved in magnitude after normalization for TIV; 2) measurements were overall more consistent than having precise agreement; 3) agreement between standardized volumes was poor in most regions; 4) compared with the output of the FS/ILP pipeline as a reference standard, the AIRC algorithm had a high specificity in flagging regional atrophy; and 5) use of the AIRC-versus-FS/

ILP output did not result in any difference in the rate of detection of neurodegenerative changes by the neuroradiologist clinicians.

Similar to the Pearson correlation, ICC estimates the strength of the relationship between 2 continuous variables. However, the Pearson correlation does not take the rater bias, which is part of the systematic error, into account. This is an important element that sets correlation apart from agreement.[22] As a result, the Pearson correlation is often paired with the intraclass correlation to optimize the detection of bias between the 2 different measurement tools. Optimized agreement requires not only a strong correlation but also low rater bias and, as a result, minimized systematic error between the 2 measurement tools. Therefore, and suggesting the presence of non-negligible bias between the 2 tools, we observed higher Pearson correlation coefficients compared with ICC-consistency and higher ICC-consistency compared with ICC-agreement for most structures, indicating the presence of rating bias among the tools (Fig 2).

Once standardized measurements were compared, the 4 main cortical lobes as well as the hippocampus demonstrated poor agreement between the FS/ILP pipeline and the AIRC tool. Because these large effect sizes are only seen in the *z* scores but not normalized volumes, they may be attributable to the differences in the composition of the normative cohort for each tool. These discrepancies might also reflect heuristic differences in the segmentation and labeling methods used by each tool. While the surface-based processing stream used by FS provides accurate delineation of white/gray matter and gray matter/CSF interfaces (Online Supplemental Data), AIRC tissue-based labeling often results in overestimation of cortical GM volumes compared with FS.[3,12,23,24] Similarly, the AIRC often undersegments and hence provides lower absolute volumes for subcortical nuclei compared with FS (Online Supplemental Data).

Participants in the AIRC normative cohort were older compared with those in the OASIS-3 group (part of the FS/ILP normative cohort, 73.25 versus 55.7 years). As a result, the normative cohort used by the AIRC might be contaminated by individuals with incipient AD pathology. This possibility is not true for OASIS-3, in which participants were followed up and remained cognitively healthy in the 3 years after the enrollment scan, on the basis of the CDR status and amyloid PET cutoffs.[5,25,26] Moreover, the AIRC normative cohort involves a relatively low number of individuals between 45 and 65 years of age, compared with OASIS-3 (approximately 20 versus 103). Because more than one-third of our participants were in this age range, the standardized score estimates made by the AIRC might be less reliable compared with those made by the FS/ILP pipeline. Given differences in normal databases, users should identify whether their patient population of interest overlaps with the age range of the normative database of any given software.

Most importantly and while FS can output both regional volumes and thicknesses, the ILP algorithm projects only percentiles calculated on the basis of regional thicknesses in the final output. Because the AIRC output is based on cortical volumes, the percentiles from the FS/ILP final report were not directly comparable with those in the AIRC report. As a result, the last step of comparing the 2 tools was to match the rate of abnormal *z* score/percentile detection on the basis of the final reports (Online Supplemental Data).

The radiologist's evaluation of volumetric brain assessments is performed on the basis of a digital report detailing the patient's *z* score/percentile for each region compared with his or her age- and sex-specific normative range. For the main lobes, this evaluation is done on the basis of cortical thicknesses from the FS/ILP versus cortical volumes from the AIRC output which might be a source of measurement bias. Not surprisingly, most of the false-positive results (8 of the 10 region/participants)- i.e. detection of abnormality in the ARIC tool in the absence of abnormal finding in the FS/ILP output- were due to thresholding differences among the tools because the FS/ILP tool has a more conservative threshold for detection of abnormalities. As a future direction we recommend a comprehensive comparison of all available FDA-cleared programs on a common neuroimaging data set, given the large number of them and that similar studies have already been performed for AD fluid biomarkers.[27-29] Finally, in developing the clinical applications of such volumetric tools additional diagnoses that were not investigated in this study, such as normal pressure hydrocephalus and primary progressive aphasia should also be considered.

While volumetric processing based on FS has been successfully used in both research and clinical settings for more than 2 decades, it lacks the time and resource efficacy in processing to permit clinical throughput in general and subspecialized radiology practices. One major driver of the long processing time and high memory usage is the reconstruction of white matter, pial, and dural surfaces, allowing FS to generate cortical thicknesses alongside cortical volumes. The AIRC output, being based on cortical volumes, has shown high sensitivity and specificity compared with the FS/ILP output, which is based on cortical volumes. On another note, rapid and accurate generation of these volumetric brain results are becoming increasingly important in high-throughput clinical settings. These features are provided by the AIRC tool due to the streamlined transfer of T1-weighted images from the PACS system, which facilitates the generation of results within several minutes and automated transfer of the results to the PACS system.

CONCLUSIONS

The AIRC brain MR tool detects volumetric changes in the main cortical lobes and subcortical regions implicated in the differential diagnosis of dementia, with sensitivity and specificity comparable with those of the FS/ILP pipeline as the reference standard. Given the much shorter processing time and streamlined user interface, the AIRC has the potential for similar comparisons in larger cohorts and further refinement of wider clinical use.

ACKNOWLEDGMENTS

We would like to thank Timothy Street and Russ Hornbeck for their critical contribution to resource management, especially software and IT support throughout this study.

Data used in this study and the normative data set used by the FS/ILP tool were in-part provided by the OASIS-4 and OASIS-3 cohorts, respectively (https://central.xnat.org/). This database in supported by the following grants: NIH P30AG066444, P50AG00561,

P30NS09857781, P01AG026276, P01AG003991, R01AG043434, UL1TR000448, R01EB009352 and P30NS098577.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Knopman DS, DeKosky ST, Cummings JL, et al. **Practice parameter: diagnosis of dementia (an evidence-based review): Report of the Quality Standards Subcommittee of the American Academy of Neurology.** *Neurology* 2001;56:1143–53 CrossRef Medline
2. Desikan RS, Ségonne F, Fischl B, et al. **An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.** *Neuroimage* 2006;31:968–80 CrossRef Medline
3. Fischl B. **FreeSurfer.** *Neuroimage* 2012;62:774–81 CrossRef Medline
4. Raji CA, Ly M, Benzinger TL. **Overview of MR imaging volumetric quantification in neurocognitive disorders.** *Top Magn Reson Imaging* 2019;28:311–15 CrossRef Medline
5. Owen CJ, Gordon B, Brier MR, et al. **The ILP: a new tool for evaluating preclinical Alzheimer's disease using volumetric MRI in a single participant.** *Alzheimers Dement* 2015;11:P697 CrossRef Medline
6. Morris JC. **The Clinical Dementia Rating (CDR).** *Neurology* 1993;43:2412–14 CrossRef Medline
7. Folstein MF, Folstein SE, McHugh PR. **"Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician.** *J Psychiatr Res* 1975;12:189–98 CrossRef Medline
8. McKhann GM, Knopman DS, Chertkow H, et al. **The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.** *Alzheimers Dement* 2011;7:263–69 CrossRef Medline
9. McKeith IG, Boeve BF, Dickson DW, et al. **Diagnosis and management of dementia with Lewy bodies: Fourth Consensus Report of the DLB Consortium.** *Neurology* 2017;89:88–100 CrossRef Medline
10. Rascovsky K, Hodges JR, Knopman D, et al. **Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia.** *Brain* 2011;134:2456–77 CrossRef Medline
11. Skrobot OA, Black SE, Chen C, et al; VICCCS group. **Progress toward standardized diagnosis of vascular cognitive impairment: guidelines from the Vascular Impairment of Cognition Classification Consensus Study.** *Alzheimers Dement* 2018;14:280–92 CrossRef Medline
12. Lyness JM, Noel TK, Cox C, et al. **Screening for depression in elderly primary care patients: a comparison of the Center for Epidemiologic Studies-Depression Scale and the Geriatric Depression Scale.** *Arch Intern Med* 1997;157:449–54 Medline
13. Schmitter D, Roche A, Maréchal B, et al; Alzheimer's Disease Neuroimaging Initiative. **An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease.** *Neuroimage Clin* 2015;7:7–17 CrossRef Medline
14. Wyman BT, Harvey DJ, Crawford K, et al; Alzheimer's Disease Neuroimaging Initiative. **Standardization of analysis sets for reporting results from ADNI MRI data.** *Alzheimers Dement* 2013;9:332–37 CrossRef Medline
15. Jack CR, Bernstein MA, Fox NC, et al; ADNI Study. **The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods.** *J Magn Reson Imaging* 2008;27:685–91 CrossRef Medline
16. LaMontagne PJ, Benzinger TLS, Morris JC, et al. **OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease.** *medRxiv* https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1. Accessed January 28, 2023
17. Koenig LN, Day GS, Salter A, et al. **Select Atrophied Regions in Alzheimer disease (SARA): an improved volumetric model for identifying Alzheimer disease dementia.** *Neuroimage Clin* 2020;26:102248 CrossRef Medline
18. Klein A, Tourville J. **101 labeled brain images and a consistent human cortical labeling protocol.** *Front Neurosci* 2012;6:171 CrossRef Medline
19. Koo TK, Li MY. **A guideline of selecting and reporting intraclass correlation coefficients for reliability research.** *J Chiropr Med* 2016;15:155–63 CrossRef Medline
20. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Routledge; 1998
21. Benjamini Y, Hochberg Y. **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society: Series B (Methodological)* 1995;57:289–300 CrossRef
22. Liu J, Tang W, Chen G, et al. **Correlation and agreement: overview and clarification of competing concepts and measures.** *Shanghai Arch Psychiatry* 2016;28:115–20 CrossRef Medline
23. Fischl B, Sereno MI, Dale AM. **Cortical surface-based analysis, II: inflation, flattening, and a surface-based coordinate system.** *Neuroimage* 1999;9:195–207 CrossRef Medline
24. Dale AM, Fischl B, Sereno MI. **Cortical surface-based analysis, I: segmentation and surface reconstruction.** *Neuroimage* 1999;9:179–94 CrossRef Medline
25. Su Y, Flores S, Hornbeck RC, et al. **Utilizing the Centiloid scale in cross-sectional and longitudinal PiB PET studies.** *Neuroimage Clin* 2018;19:406–16 CrossRef Medline
26. Su Y, Flores S, Wang G, et al. **Comparison of Pittsburgh compound B and florbetapir in cross-sectional and longitudinal studies.** *Alzheimers Dement (Amst)* 2019;11:180–90 CrossRef Medline
27. Janelidze S, Teunissen CE, Zetterberg H, et al. **Head-to-head comparison of 8 plasma amyloid-β 42/40 assays in Alzheimer disease.** *JAMA Neurol* 2021;78:1375–82 CrossRef Medline
28. Pemberton HG, Goodkin O, Prados F, et al; Alzheimer's Disease Neuroimaging Initiative. **Automated quantitative MRI volumetry reports support diagnostic interpretation in dementia: a multi-rater, clinical accuracy study.** *Eur Radiol* 2021;31:5312–23 CrossRef Medline
29. Pemberton HG, Zaki LA, Goodkin O, et al. **Technical and clinical validation of commercial automated volumetric MRI tools for dementia diagnosis: a systematic review.** *Neuroradiology* 2021;63:1773–89 CrossRef Medline