HHS Public Access

Multi-Site Cross-Site Inter-Rater and Test-Retest Reliability and Construct Validity of the MarkVCID White Matter Hyperintensity Growth and Regression Protocol

Ahmed A. Bahrani[a,b,*], **Erin L. Abner**[b,c], **Charles S. DeCarli**[d], **Justin M. Barber**[b], **Abigail C. Sutton**[b], **Pauline Maillard**[d], **Francisco Sandoval**[d], **Konstantinos Arfanakis**[e,f], **Yung-Chuan Yang**[e], **Arnold M. Evia**[e], **Julie A. Schneider**[e,g], **Mohamad Habes**[h,i], **Crystal G. Franklin**[h], **Sudha Seshadri**[i], **Claudia L. Satizabal**[i], **Arvind Caprihan**[j], **Jeffrey F. Thompson**[j], **Gary A. Rosenberg**[k], **Danny J.J. Wang**[l], **Kay Jann**[l], **Chenyang Zhao**[l], **Hanzhang Lu**[m], **Paul B. Rosenberg**[n], **Marilyn S. Albert**[o], **Doaa G. Ali**[b], **Herpreet Singh**[p], **Kristin Schwab**[p], **Steven M. Greenberg**[p], **Karl G. Helmer**[q], **David K. Powel**[r,s], **Brian T. Gold**[b,r,s], **Larry B. Goldstein**[a], **Donna M. Wilcock**[b,t], **Gregory A. Jicha**[a,b]

[a]Department of Neurology, University of Kentucky, College of Medicine, Lexington, KY, USA

[b]Sanders-Brown Center on Aging, University of Kentucky, College of Medicine, Lexington, KY, USA

[c]Department of Epidemiology & Environmental Health, University of Kentucky, College of Public Health, Lexington, KY, USA

[d]Department of Neurology, University of California, Davis, CA, USA

[e]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA

[f]Department of Biomedical Engineering, Illinois Institute of Technology, Chicago, IL, USA

[g]Department of Neurological Sciences, Rush University Medical Center, Chicago, IL, USA

[h]Research Imaging Institute, University of Texas Health San Antonio, San Antonio, TX, USA

[i]Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, University of Texas Health San Antonio, San Antonio, TX, USA

[j]The Mind Research Network, Albuquerque, NM, USA

[k]Center for Memory and Aging, University of New Mexico, Health Sciences Center, Albuquerque, NM, USA

[l]Departments of Neurology and Radiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[*]Correspondence to: Ahmed A. Bahrani, PhD, Department of Neurology, College of Medicine, University of Kentucky, 740 Rose St., USA. ahmed.bahrani@uky.edu.

[m]Department of Radiology and Radiological Science, Johns Hopkins University, School of Medicine, Baltimore, MD, USA

[n]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, School of Medicine, Baltimore, MD, USA

[o]Department of Neurology, Johns Hopkins University, School of Medicine, Baltimore, MD, USA

[p]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

[q]Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

[r]Department of Neuroscience, University of Kentucky, College of Medicine, Lexington, KY, USA

[s]Magnetic Resonance Imaging and Spectroscopy Center, University of Kentucky, Lexington, KY, USA

[t]Department of Physiology, University of Kentucky, College of Medicine, Lexington, KY, USA

Abstract

Background: White matter hyperintensities (WMH) that occur in the setting of vascular cognitive impairment and dementia (VCID) may be dynamic increasing or decreasing volumes or stable over time. Quantifying such changes may prove useful as a biomarker for clinical trials designed to address vascular cognitive-impairment and dementia and Alzheimer's Disease.

Objective: Conducting multi-site cross-site inter-rater and test-retest reliability of the MarkVCID white matter hyperintensity growth and regression protocol.

Methods: The NINDS-supported MarkVCID Consortium evaluated a neuroimaging biomarker developed to track WMH change. Test-retest and cross-site inter-rater reliability of the protocol were assessed. Cognitive test scores were analyzed in relation to WMH changes to explore its construct validity.

Results: ICC values for test-retest reliability of WMH growth and regression were 0.969 and 0.937 respectively, while for cross-site inter-rater ICC values for WMH growth and regression were 0.995 and 0.990 respectively. Word list long-delay free-recall was negatively associated with WMH growth ($p < 0.028$) but was not associated with WMH regression.

Conclusions: The present data demonstrate robust ICC validity of a WMH growth/regression protocol over a one-year period as measured by cross-site inter-rater and test-retest reliability. These data suggest that this approach may serve an important role in clinical trials of disease-modifying agents for VCID that may preferentially affect WMH growth, stability, or regression.

Keywords

INTRODUCTION

White matter hyperintensities (WMH) found on magnetic resonance imaging (MRI) brain scans are a widely-recognized imaging correlate for the assessment of small vessel-type cerebrovascular disease (CVD) burden in vascular cognitive impairment and dementia (VCID) and Alzheimer's disease (AD) in both cohort and population studies [1–3]. WMH in late life are associated with small vessel ischemic disease and vascular risk factors such as hypertension, hyperlipidemia, and diabetes, among others [4–7].

WMH are dynamic, often increasing in volume over time, but can also be stable and even regress [8–12]. Despite a limited understanding of specific risk factor contributions to WMH and the longitudinal nature of such lesions, WMH volumes are associated with deficits in cognitive function, including both memory and executive function [13–15]. As such, WMH volumes represent an important potential biomarker for assessing WMH-related VCID as a part of entry criteria for large scale multi-site clinical trials for vascular dementia and AD dementia, as well as a potential surrogate outcome of treatment mechanisms. Measuring such changes in WMH over periods as short as one year is critical for both mechanistic discoveries and for the informed future use of biomarkers to track disease modification in clinical trials of VCID and AD.

Although several studies measured and/or predicted WMH change over time, these have focused on total WMH volume measures rather than assessing regional WMH changes [12, 16–19], and/or lacked data showing that the protocol could be performed reliably across MRI analysts and between centers [20–23]. In this setting, total WMH volume changes over time have been defined as the arithmetic subtraction of summed whole brain WMH volumes across two or more time points [10, 12]. Such methods assume that only global WMH burden is important, which ignores the possibility that mixed WMH changes (i.e., growth, regression, and/or stable volumes) may occur concomitantly within a given individual in different anatomic regions and or even within discrete WMHs.

Understanding the relative contributions of such dynamic changes in the progression of VCID and how discrete growth, regression, and stability of WMH lesions may be altered by specific risk factors or therapeutic interventions may provide valuable insights into both potential mechanisms of disease and future therapeutic interventions. Major obstacles to the development of dynamic WMH protocols include difficulties in co-registration of fluid attenuated inversion recovery (FLAIR) images, which is often complicated by structural volume changes and fluctuations in WMH intensity between image acquisitions, despite using standardized acquisition protocols.

We developed a protocol that provides a reliable measure of WMH changes over periods as short as one-year [24]. This protocol was assessed previously for single-site inter-rater reliability (i.e., it was tested by two independent raters at one site), but not for cross-site inter-rater or test-retest reliability. In the current study, the NINDS-supported MarkVCID Consortium tested this protocol for cross-site inter-rater and test-retest reliability. In addition, we tested the hypothesis that measurement of discrete longitudinal change in WMH has a meaningful clinical correlation (i.e., construct validity).

METHODS

Study design

The MarkVCID approach for assessing MRI-based candidate biomarkers has been reported previously [25]. Inter-rater assessments used a standardized set of MRI acquisitions at baseline and one-year follow-up that were processed by independent raters across six sites of the MarkVCID consortium.

Test-retest reliability assessment involved comparisons of discrete WMH growth (i.e., conversion of normal appearing WM (NAWM) to WMH) and regression (i.e., conversion of WMH to NAWM) measures between each of two baseline scans, obtained 7–14 days apart, in relation to a single one-year follow-up scan for each participant. Figure 1A is a schematic of the study. Briefly, the test-retest reliability assessment was performed in two steps. First, each of MarkVCID sites acquired two scans (7–14 days between the first and second scans) and a single scan at one-year follow-up visit. Then, the test-retest reliability analysis was completely performed at a single site (University of Kentucky) utilizing the scans from the MarkVCID sites including the scans from the University of Kentucky.

An exploratory investigation of construct validity was performed in a subset of cross-site MarkVCID participants in relation to supraspan word list long delayed free recall (LDFR) [26] based on prior data demonstrating that supraspan word list delayed free recall was correlated with total WMH volume changes over a one-year interval [12, 27].

Participants

Participants, comprising persons with low and high CVD risk, with WMH Fazekas scores 1 at baseline were recruited into the MarkVCID cohort across the MarkVCID sites. The complete inclusion/exclusion criteria for the MarkVCID participants have been published [28]. In general, each site followed specific recruitment criteria to create a generalizable MarkVCID validation data set including cognitive status, vascular risk factors, age, gender, etc., that could be used for future studies. MarkVCID was approved by the Institutional Review Boards (IRBs) at each participating site and all participants provided informed consent.

For the test-retest reliability analysis, Data from 18 participants who completed test-retest and follow-up visits were identified for test-retest assessment. Data from 30 participants were randomly selected from the group of the initial participants ($n = 97$ by May 2022) completing all baseline and one-year follow-up clinical and imaging acquisitions for inter-rater reliability assessments as specified in the a priori analysis plan for this study. Prior to reliability assessments, all cases were reviewed by the University of Kentucky as the kit leading site to ensure an adequate distribution of clinical and WMH volume scores and measurements and to exclude any inconsistent scans for example with motion artifacts or any possible changes to the scan sequence parameters prior to the follow-up scans. The selection of cases was otherwise arbitrary regarding participant characteristics and scanner vendors and models including specific acquisition parameters (which vary based on scanner type), and site staff characteristics.

For the exploratory investigation of clinical correlates, Data of additional group of participants with longitudinal imaging and clinical data from the MarkVCID consortium were also collected ($n = 53$). The imaging and clinical data from these participants were used for exploratory evaluation of the construct validity of the WMH growth/regression protocol.

Image acquisition

Multi-echo 3D T1-weighted and 3D FLAIR MRI sequences [25] were acquired longitudinally for the WMH growth/regression protocol at the baseline and 1-year follow-up visits. Participants at all sites were scanned using a harmonized protocol that was optimized to generate similar image contrast irrespective of scanner type [25]. Scanners were from two different vendors and three different models (Siemens, TRIO & PRISMA and Philips ACHIEVA), but all had a magnetic field strength of 3.0T [25]. The image resolutions for both FLAIR and 3D T1-weighted images were $1 \times 1 \times 1$ mm, and the field of view was 256×256 mm. Bandwidth was matched as closely as possible to match the image distortions to increase the accuracy of the image registration between the two scan types. All sites were asked to use the same head coil for all scans for each participant to ensure identical scans for each individual site. Also, instructions were given to each site for reducing head movement, consistency of head coil placement, and consistency in selecting the field of view in an identical manner for each time point.

WMH growth/regression longitudinal protocol

The details of the longitudinal protocol have been published previously [24]. Briefly, the longitudinal protocol was based on quantitative WMH protocols that were developed at the University of Kentucky's Sanders-Brown Center on Aging [29, 30]. 3D anatomical T1-weighted and 3D FLAIR images were acquired for each participant at baseline and at the one-year follow-up visit. Intensity correction (N3-correction) was performed for all images to correct the intensity non-uniformity [31–33]. The T1-weighted image was registered to the FLAIR image, using a 6-parameter rigid-body registration, for each time point utilizing SPM12 (http://www.fil.ion.ucl.ac.uk/spm). The resultant two time-point FLAIR images were registered again using SPM12 longitudinal registration tool to generate a midpoint image and deformation field maps for each time-point to be used to register the two time-point images to the midpoint image as previously described [24]. Images with motion artifacts were excluded from the study. The midpoint images were skull-stripped using the FSL-BET FMRIB software library (FSL v5.0.8) Brain Extraction Tool (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET) and segmented utilizing the SPM12 unified segmentation tool to generate a WM mask [24, 30, 34]. The histogram distribution curve for each FLAIR WM image was fitted to the Gaussian model curve to compute the maximum and minimum threshold values to generate the WMH mask [24]. The processing pipeline was scripted within a Singularity container (https://singularity.lbl.gov) for consistent usage across sites.

As any WMH mask may include artifacts (false-positive voxels) due to the intensity overlaps between the normal-appearing WM and the WMH, a manual editing step was performed independently by all raters [29]. All sites received uniform training in manual editing procedures [29]. Using the same protocol, scripts within the Singularity container allowed

the generation of binary growth and regression WMH masks based on a simple subtraction between the total WMH masks of the two-time points. Positive voxels represent WMH growth, and negative voxels represent WMH regression (Fig. 1B).

Data analysis

Reliability—Test-retest reliability analyses were performed at a single site using images (n = 18) acquired across the MarkVCID consortium sites using standard ICC analysis that included a two-way (raters and participants) mixed effects model [35]. For the ICC analyses, rho 0.85 [36, 37] was set as the minimally acceptable agreement for test-retest and cross-site inter-rater reliability. The Levene Statistic (test of Homogeneity of Variances) was used to test the homogeneity of variances across sites (p-values<0.05 were required to reject the null hypothesis).

Exploratory clinical construct validity

To evaluate construct validity, a preliminary analysis on a small sample of MarkVCID participants (n = 53) was conducted using an *a priori* statistical analysis plan that used the supraspan word list LDFR as the clinical correlate. For these analyses, WMH growth and regression volumes were normalized to intracranial volume, and then log-transformed to improve the plausibility of the linear model assumptions.

Raw LDFR scores were downloaded from MarkVCID1 database. The MarkVCID protocol allowed sites to use their choice of California Verbal Learning Test-second edition (CVLT-II; target words = 16), Hopkins Verbal Learning Test (HVLT; 12), and CVLT-II Short Form (CVLT-II-SF; 9). Due to the word lists differing in maximum target words, all raw LDFR scores were standardized to z-scores using published reference data for the respective word lists. For the CVLT-II and CVLT-II-SF we used the lookup tables in Delis et al. (2000) [38]. For the HVLT, we used the means and SDs from Ryan et al. (2021) using Table 3 [39]. The form version for CVLT-II (i.e., Standard or Alternate) is not captured in the MarkVCID1 database so we used the Standard Form reference tables from Delis et al. (2000) for all instances of CVLT-II. The form version for HVLT is also not captured in the MarkVCID1 database, but Ryan et al. (2021) used Form 6 in developing their norms [39]. The CVLT-II and CVLT-II-SF reference norms account for age at visit and sex. The HVLT reference norms account for sex, education, ethno-racial group, and age at visit. Despite these discrepancies in the segmentation of reference norms for the HVLT compared to CVLT-II and CVLT-II-SF, all reference data were derived from large, nationally representative samples. The standardized LDFR data at 12-month follow-up were regressed on measures of WMH change (growth and regression), controlling for the LDFR z-scores at the baseline.

RESULTS

Participant demographics, including age, education, and sex at birth, as well as general clinical status measured using the Montreal Cognitive Assessment test scores are provided in Table 1.

Test-retest reliability

Test-retest ICCs of the two-way mixed consistency and absolute agreement models for WMH growth were identical at 0.969 (95% confidence interval CIrange 0.919–0.988). The two-way mixed consistency ICC was 0.939 (95% CI range 0.838–0.977) for WMH regression test-retest, and the two-way mixed absolute agreement model was 0.937 (95% CI range 0.835–0.976; Fig. 2).

Cross-site inter-rater reliability

The mean ICC for growth was 0.995 (CI of 0.991–0.997), and the mean ICC for regression was 0.990 (CI of 0.984–0.995; Fig. 3). Moreover, the ICC of the WMH masks before editing was tested to check the reliability of the kit fully automated part and the mean ICC value was 1 at each time point. The Levene Statistic test provided no evidence of heterogeneity across sites. Levene coefficients for WMH (baseline), WMH (follow-up), WMH growth and WMH regression were 0.003 ($p = 1$), 0.007 ($p = 1$), 0.094 ($p = 0.997$), and 0.183 ($p = 0.981$), respectively. Thus, these results provided no evidence of heterogeneity across-sites, supporting the ICC test for cross-site inter-rater reliability.

Variability in WMH growth, stability, and regression over a one-year period

Figure 4 shows participant-specific WMH volume changes (increasing (growth), decreasing (regression), and stability) over a one-year follow-up for the 53 participants included in the analysis of construct validity. Although WMH volume stability is the dominant result over the year of follow-up, within-participant WMH volume changes were notable and varied significantly among participants irrespective of total WMH change.

Preliminary analysis of construct validity

Adjusted linear regression analyses demonstrated that a 1-unit increase in log WMH growth was associated with about a 75% of standard deviation lower mean LDFR z-score at follow-up (95% confidence interval: −1.39, −0.093; $p = 0.028$); model $R^2 = 0.59$ (Fig. 5). In contrast, WMH regression was not significantly associated with mean LDFR z-score ($p = 0.54$), but the association was in the same direction (about 25% of standard deviation lower z-score with a 1-unit increase in log WMH regression.

DISCUSSION

In this study, we evaluated test-retest and cross-site inter-rater reliability of a novel neuroimaging biomarker that measures WMH growth, regression, and stability. These data, obtained from the MarkVCID consortium, demonstrated strong reliability for the assessment of longitudinal WMH change over one-year and may be a useful biomarker for future interventional studies of VCID, AD and for other targeted diseases and interventions in which the burden of WM lesions is important. Exploratory analysis using WMH growth and regression to predict memory performance demonstrated that isolating WMH growth may be a more powerful for clinical investigation than relying on the global WMH difference volume that has mixed effect of WMH growth and regression.

Advancing clinical trials for VCID is a research priority that is in part dependent on the development of pharmacodynamic biomarkers, which are currently lacking [40–42]. Although whole-brain WMH, DTI, and ASL changes applied cross-sectionally as repeated measures show some promise, these imaging features are highly dependent on a complex interplay of both tissue destructive and healing/reparative factors. A biomarker based on WMH dynamic changes may provide a window into new processes in the development and evolution of VCID.

Optimal management of CVD risk factors, increased angiogenesis, and reductions in central nervous system inflammation may facilitate WMH regression, whereas poorly controlled CVD risk factors, reduced angiogenesis, and increased central nervous system inflammation may preferentially increase WMH growth. Evaluating WMH growth and regression using this protocol or similar methods may allow investigators to choose the VCID cohort (inclusion/exclusion) criteria most likely to be selectively in need of reduced WMH growth, stability and or augmented regression, most likely to be influenced by a candidate intervention. Such a discovery could yield insights into basic mechanisms responsible for VCID that are essential for progress in the field. Therefore, this protocol could be used to track the effectiveness of an intervention, pharmaceutical, etc., on the progression of a disease. For example, the SPRINT MIND trial found that lowering systolic blood pressure to less than 120 mmHg was associated with a reduction in the growth of WMH [43]. It remains unclear if such aggressive blood pressure management may also augment regression of WMH, and thus, whether WMH regression can serve as a marker of clinical benefit. Further understanding of these relationships, using refined biomarkers for VCID, is critical for optimal scientific discovery into disease mechanisms and interpretation of data stemming from interventional trials.

Limitations of the current study include the use of a participant group that is largely White/ Caucasian, highly educated, and that has agreed to participate in an observational study for biomarker development in VCID. Further evaluation in diverse cohorts that can assess the generalizability of the current findings is essential. Although tested in a relatively small cohort of patients, the robustness of our findings calls into question the need for larger reliability studies. It should also be recognized that the MarkVCID consortium of sites represent a faculty that is expert in the assessment of biomarkers of VCID and it remains unclear how robust the protocol might be in settings with less expertise in the assessment of VCID biomarkers. Further work in centers that are not part of the MarkVCID consortium is needed to assess the reliability and validity of widespread use of the WMH growth and regression protocol. The Singularity Container for this protocol has been designed to advance knowledge of WMH longitudinal changes irrespective of the specific expertise or skills of the faculty and or staff performing such analyses at individual sites.

The need for a manual editing step can also be considered a limitation of the present protocol. The requirement for final manual editing is a critical issue for use of this protocol as the artifacts inherent in 3D FLAIR acquisition and post-processing steps are substantial. All site staff across all centers engaged in our study had specific training in the use of the protocol with a focus on the final manual editing step. The data presented attest to the success of such training and the usability of the MarkVCID WMH growth and

regression protocol. Ongoing work seeks to fully automate the protocol using machine learning algorithms to bypass this manual process, but until such time, human analysis remains an essential part of the process. Indeed, data such as that derived from the current protocol may well serve as the "gold-standard" for machine learning algorithms that require such data to refine more automated processes.

In summary, despite the caveats presented above, the present data demonstrate that the MarkVCID WMH Growth/Regression protocol is reliable for use as a potentially valuable tool for future discoveries and as an outcome measure for clinical trials targeting VCID and AD in an aging population. Therefore, the protocol can be used to process the data despite the brand of the MRI scanner; however, for group comparison, it is important for the scan parameters to be as close as possible for different MRI scanners. Further biologic and clinical validation of this protocol is currently in progress.

ACKNOWLEDGMENTS

DATA AVAILABILITY

The data are not publicly available. To request a copy of the raw data, please contact the MarkVCID consortium.

REFERENCES

[1]. Meyer JS, Kawamura J, Terayama Y (1992) White matter lesions in the elderly. J Neurol Sci 110, 1–7. [PubMed: 1506848]

[2]. Gootjes L, Teipel SJ, Zebuhr Y, Schwarz R, Leinsinger G, Scheltens P, Moller HJ, Hampel H (2004) Regional distribution of white matter hyperintensities in vascular dementia, Alzheimer's disease and healthy aging. Dement Geriatr Cogn Disord 18, 180–188. [PubMed: 15211074]

[3]. Debette S, Markus HS (2010) The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. BMJ 341, c3666. [PubMed: 20660506]

[4]. Kalaria RN, Erkinjuntti T (2006) Small vessel disease and subcortical vascular dementia. J Clin Neurol 2, 1–11. [PubMed: 20396480]

[5]. Chowdhury MH, Nagai A, Bokura H, Nakamura E, Kobayashi S, Yamaguchi S (2011) Age-related changes in white matter lesions, hippocampal atrophy, and cerebral microbleeds in healthy subjects without major cerebrovascular risk factors. J Stroke Cerebrovasc Dis 20, 302–309. [PubMed: 20634092]

[6]. Duan D, Shen L, Cui C, Shu T, Zheng J (2017) Association between Low-density lipoprotein cholesterol and occipital periventricular hyperintensities in a group of Chinese patients: An observational study. Lipids Health Dis 16, 48. [PubMed: 28241772]

[7]. Jimenez-Conde J, Biffi A, Rahman R, Kanakis A, Butler C, Sonni S, Massasa E, Cloonan L, Gilson A, Capozzo K, Cortellini L, Ois A, Cuadrado-Godia E, Rodriguez-Campello A, Furie KL, Roquer J, Rosand J, Rost NS (2010) Hyperlipidemia and reduced white matter hyperintensity volume in patients with ischemic stroke. Stroke 41, 437–442. [PubMed: 20133919]

[8]. Ramirez J, McNeely AA, Berezuk C, Gao F, Black SE (2016) Dynamic progression of white matter hyperintensities in Alzheimer's disease and normal aging: Results from the Sunnybrook Dementia Study. Front Aging Neurosci 8, 62. [PubMed: 27047377]

[9]. Silbert LC, Nelson C, Howieson DB, Moore MM, Kaye JA (2008) Impact of white matter hyperintensity volume progression on rate of cognitive and motor decline. Neurology 71, 108–113. [PubMed: 18606964]

[10]. van Leijsen EM, Bergkamp MI, van Uden IW, Cooijmans S, Ghafoorian M, van der Holst HM, Norris DG, Kessels RP, Platel B, Tuladhar AM, de Leeuw FE (2019) Cognitive consequences of regression of cerebral small vessel disease. Eur Stroke J 4, 85–89. [PubMed: 31165098]

[11]. van Leijsen EMC, van Uden IWM, Ghafoorian M, Bergkamp MI, Lohner V, Kooijmans ECM, van der Holst HM, Tuladhar AM, Norris DG, van Dijk EJ, Rutten-Jacobs LCA, Platel B, Klijn CJM, de Leeuw FE (2017) Nonlinear temporal dynamics of cerebral small vessel disease: The RUN DMC study. Neurology 89, 1569–1577. [PubMed: 28878046]

[12]. Al-Janabi OM, Bauer CE, Goldstein LB, Murphy RR, Bahrani AA, Smith CD, Wilcock DM, Gold BT, Jicha GA (2019) White matter hyperintensity regression: Comparison of brain atrophy and cognitive profiles with progression and stable groups. Brain Sci 9, 170. [PubMed: 31330933]

[13]. Brickman AM, Muraskin J, Zimmerman ME (2009) Structural neuroimaging in Alzheimer's disease: Do white matter hyperintensities matter? Dialogues Clin Neurosci 11, 181–190. [PubMed: 19585953]

[14]. Iorio M, Spalletta G, Chiapponi C, Luccichenti G, Cacciari C, Orfei MD, Caltagirone C, Piras F (2013) White matter hyperintensities segmentation: A new semi-automated method. Front Aging Neurosci 5, 76. [PubMed: 24339815]

[15]. Tsai JZ, Peng SJ, Chen YW, Wang KW, Li CH, Wang JY, Chen CJ, Lin HJ, Smith EE, Wu HK, Sung SF, Yeh PS, Hsin YL (2014) Automated segmentation and quantification of white matter hyperintensities in acute ischemic stroke patients with cerebral infarction. PLoS One 9, e104011. [PubMed: 25127120]

[16]. Carmichael O, Schwarz C, Drucker D, Fletcher E, Harvey D, Beckett L, Jack CR Jr., Weiner M, DeCarli C (2010) Longitudinal changes in white matter disease and cognition in the first year of the Alzheimer disease neuroimaging initiative. Arch Neurol 67, 1370–1378. [PubMed: 21060014]

[17]. Chen SP, Chou KH, Fuh JL, Huang YH, Huang CC, Lirng JF, Wang YF, Lin CP, Wang SJ (2018) Dynamic changes in white matter hyperintensities in reversible cerebral vasoconstriction syndrome. JAMA Neurol 75, 1106–1113. [PubMed: 29868878]

[18]. Kim SJ, Lee DK, Jang YK, Jang H, Kim SE, Cho SH, Kim JP, Jung YH, Kim EJ, Na DL, Lee JM, Seo SW, Kim HJ (2020) The effects of longitudinal white matter hyperintensity change on cognitive decline and cortical thinning over three years. J Clin Med 9, 2663. [PubMed: 32824599]

[19]. van den Heuvel DM, ten Dam VH, de Craen AJ, Admiraal-Behloul F, van Es AC, Palm WM, Spilt A, Bollen EL, Blauw GJ, Launer L, Westendorp RG, van Buchem MA (2006) Measuring longitudinal white matter changes: Comparison of a visual rating scale with a volumetric measurement. AJNR Am J Neuroradiol 27, 875–878. [PubMed: 16611781]

[20]. Pozorski V, Oh JM, Okonkwo O, Krislov S, Barzgari A, Theisen F, Sojkova J, Bendlin BB, Johnson SC, Gallagher CL (2019) Cross-sectional and longitudinal associations between total and regional white matter hyperintensity volume and cognitive and motor function in Parkinson's disease. Neuroimage Clin 23, 101870. [PubMed: 31150958]

[21]. Shao Y, Chen Z, Ming S, Ye Q, Shu Z, Gong C, Pang P, Gong X (2018) Predicting the development of normal-appearing white matter with radiomics in the aging brain: A longitudinal clinical study. Front Aging Neurosci 10, 393. [PubMed: 30546304]

[22]. Maillard P, Carmichael O, Fletcher E, Reed B, Mungas D, DeCarli C (2012) Coevolution of white matter hyperintensities and cognition in the elderly. Neurology 79, 442–448. [PubMed: 22815562]

[23]. Maillard P, Fletcher E, Lockhart SN, Roach AE, Reed B, Mungas D, DeCarli C, Carmichael OT (2014) White matter hyperintensities and their penumbra lie along a continuum of injury in the aging brain. Stroke 45, 1721–1726. [PubMed: 24781079]

[24]. Bahrani AA, Smith CD, Barber JM, Al-Janabi OM, Powell DK, Andersen AH, Ramey BD, Abner EL, Goldstein LB, Winder Z, Gold BT, Van Eldik L, Wilcock DM, Jicha GA (2021) Development of a protocol to assess within-subject, regional white matter hyperintensity changes in aging and dementia. J Neurosci Methods 360, 109270. [PubMed: 34171312]

[25]. Lu H, Kashani AH, Arfanakis K, Caprihan A, DeCarli C, Gold BT, Li Y, Maillard P, Satizabal CL, Stables L, Wang DJJ, Corriveau RA, Singh H, Smith EE, Fischl B, van der Kouwe A, Schwab K, Helmer KG, Greenberg SM (2021) MarkVCID cerebral small vessel consortium: II. Neuroimaging protocols. Alzheimers Dement 17, 716–725. [PubMed: 33480157]

[26]. Crosson B, Novack TA, Trenerry MR, Craig PL (1988) California Verbal Learning Test (CVLT) performance in severely head-injured and neurologically normal adult males. J Clin Exp Neuropsychol 10, 754–768. [PubMed: 3235649]

[27]. Gifford KA, Liu D, Neal JE, Babicz MA, Thompson JL, Walljasper LE, Wiggins ME, Turchan M, Pechman KR, Osborn KE, Acosta LMY, Bell SP, Hohman TJ, Libon DJ, Blennow K, Zetterberg H, Jefferson AL (2018) The 12-Word Philadelphia Verbal Learning Test performances in older adults: Brain MRI and cerebrospinal fluid correlates and regression-based normative data. Dement Geriatr Cogn Dis Extra 8, 476–491. [PubMed: 30631339]

[28]. Wilcock D, Jicha G, Blacker D, Albert MS, D'Orazio LM, Elahi FM, Fornage M, Hinman JD, Knoefel J, Kramer J, Kryscio RJ, Lamar M, Moghekar A, Prestopnik J, Ringman JM, Rosenberg G, Sagare A, Satizabal CL, Schneider J, Seshadri S, Sur S, Tracy RP, Yasar S, Williams V, Singh H, Mazina L, Helmer KG, Corriveau RA, Schwab K, Kivisakk P, Greenberg SM (2021) MarkVCID cerebral small vessel consortium: I. Enrollment, clinical, fluid protocols. Alzheimers Dement 17, 704–715. [PubMed: 33480172]

[29]. Bahrani AA, Al-Janabi OM, Abner EL, Bardach SH, Kryscio RJ, Wilcock DM, Smith CD, Jicha GA (2019) Post-acquisition processing confounds in brain volumetric quantification of white matter hyperintensities. J Neurosci Methods 327, 108391. [PubMed: 31408649]

[30]. Bahrani AA, Powell DK, Yu G, Johnson ES, Jicha GA, Smith CD (2017) White matter hyperintensity associations with cerebral blood flow in elderly subjects stratified by cerebrovascular risk. J Stroke Cerebrovasc Dis 26, 779–786. [PubMed: 28063772]

[31]. Sled JG, Zijdenbos AP, Evans AC (1998) A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17, 87–97. [PubMed: 9617910]

[32]. Boyes RG, Gunter JL, Frost C, Janke AL, Yeatman T, Hill DL, Bernstein MA, Thompson PM, Weiner MW, Schuff N, Alexander GE, Killiany RJ, DeCarli C, Jack CR, Fox NC (2008) Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. Neuroimage 39, 1752–1762. [PubMed: 18063391]

[33]. Ashburner J, Friston KJ (2005) Unified segmentation. Neuroimage 26, 839–851. [PubMed: 15955494]

[34]. Smith CD, Johnson ES, Van Eldik LJ, Jicha GA, Schmitt FA, Nelson PT, Kryscio RJ, Murphy RR, Wellnitz CV (2016) Peripheral (deep) but not periventricular MRI white matter hyperintensities are increased in clinical vascular dementia compared to Alzheimer's disease. Brain Behav 6, e00438. [PubMed: 26925303]

[35]. Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86, 420–428. [PubMed: 18839484]

[36]. Cicchetti DV (1994) Multiple comparison methods: Establishing guidelines for their valid application in neuropsychological research. J Clin Exp Neuropsychol 16, 155–161. [PubMed: 8150886]

[37]. Cicchetti DV, Showalter D (1997) A computer program for assessing interexaminer agreement when multiple ratings are made on a single subject. Psychiatry Res 72, 65–68. [PubMed: 9355820]

[38]. Delis DC, Hramer JH, Kaplan E, Ober BA (2000) California verbal learning test, Pearson.

[39]. Ryan J, Woods RL, Murray AM, Shah RC, Britt CJ, Reid CM, Wolfe R, Nelson MR, Lockery JE, Orchard SG, Trevaks RE, Chong TJ, McNeil JJ, Storey E (2021) Normative performance of older

individuals on the Hopkins Verbal Learning Test-Revised (HVLT-R) according to ethno-racial group, gender, age and education level. Clin Neuropsychol 35, 1174–1190. [PubMed: 32100619]

[40]. Cummings J (2019) The role of biomarkers in Alzheimer's disease drug development. Adv Exp Med Biol 1118, 29–61. [PubMed: 30747416]

[41]. Friedman LG, McKeehan N, Hara Y, Cummings JL, Matthews DC, Zhu J, Mohs RC, Wang D, Hendrix SB, Quintana M, Schneider LS, Grundman M, Dickson SP, Feldman HH, Jaeger J, Finger EC, Ryan JM, Niehoff D, Dickinson SL, Markowitz JT, Owen M, Travaglia A, Fillit HM (2021) Value-generating exploratory trials in neurodegenerative dementias. Neurology 96, 944–954. [PubMed: 33674360]

[42]. Sweeney P, Park H, Baumann M, Dunlop J, Frydman J, Kopito R, McCampbell A, Leblanc G, Venkateswaran A, Nurmi A, Hodgson R (2017) Protein misfolding in neurodegenerative diseases: Implications and strategies. Transl Neurodegener 6, 6. [PubMed: 28293421]

[43]. Williamson JD, Pajewski NM, Auchus AP, Bryan RN, Chelune G, Cheung AK, Cleveland ML, Coker LH, Crowe MG, Cushman WC, Cutler JA, Davatzikos C, Desiderio L, Erus G, Fine LJ, Gaussoin SA, Harris D, Hsieh MK, Johnson KC, Kimmel PL, Tamura MK, Launer LJ, Lerner AJ, Lewis CE, Martindale-Adams J, Moy CS, Nasrallah IM, Nichols LO, Oparil S, Ogrocki PK, Rahman M, Rapp SR, Reboussin DM, Rocco MV, Sachs BC, Sink KM, Still CH, Supiano MA, Snyder JK, Wadley VG, Walker J, Weiner DE, Whelton PK, Wilson VM, Woolard N, Wright JT Jr., Wright CB (2019) Effect of intensive vs standard blood pressure control on probable dementia: A randomized clinical trial. JAMA 321, 553–561. [PubMed: 30688979]
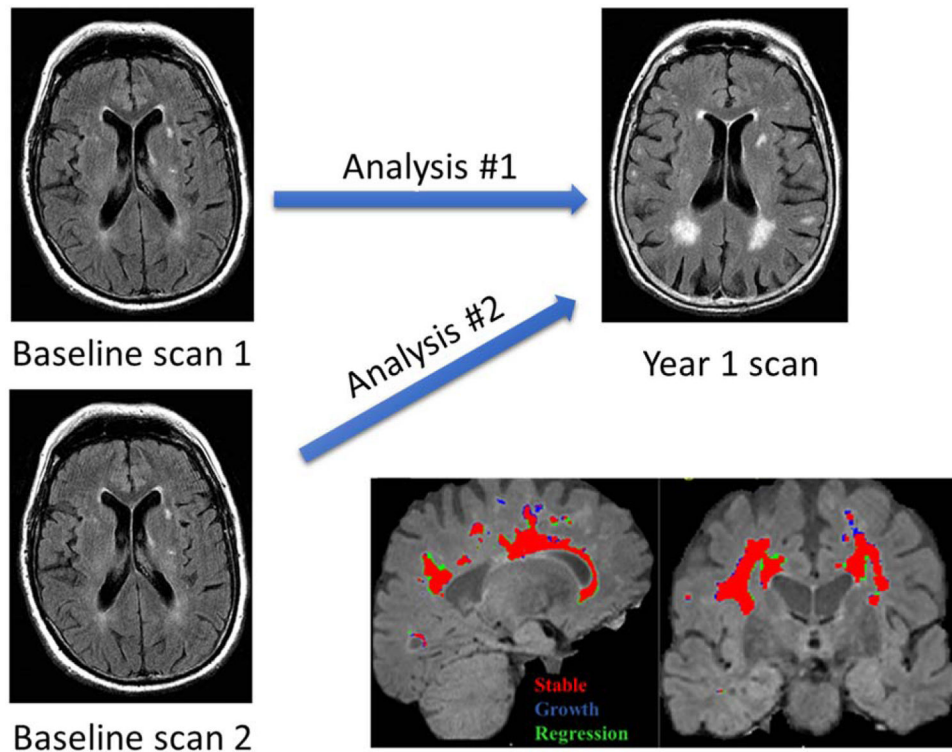
Fig. 1.
Strategy for determining test-retest repeatability for this longitudinal protocol and the source of image acquisition variability. A) This figure schematically illustrates the test-retest study design that utilized a comparison of two unique baseline scans with a single follow-up visit scan. B) Dynamic WMH changes within an individual subject are shown, (Red: Stable, Blue: Growth, Green: Regression).
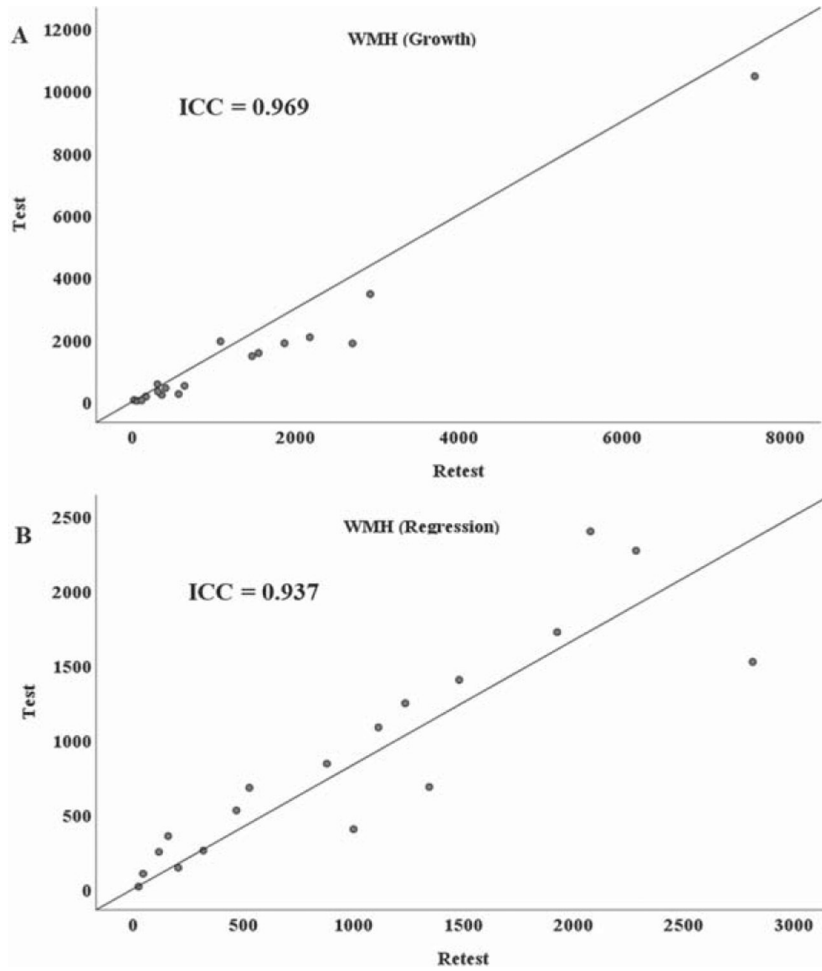
Fig. 2.
Test-Retest Repeatability scatter plots for WMH A) Growth, and B) Regression volumes in mm^3. The figure demonstrates highly significant ICC values ($p < 0.001$) for both WMH growth and regression comparing the two repeat baseline scans (test and retest) to the single one-year follow-up scan.
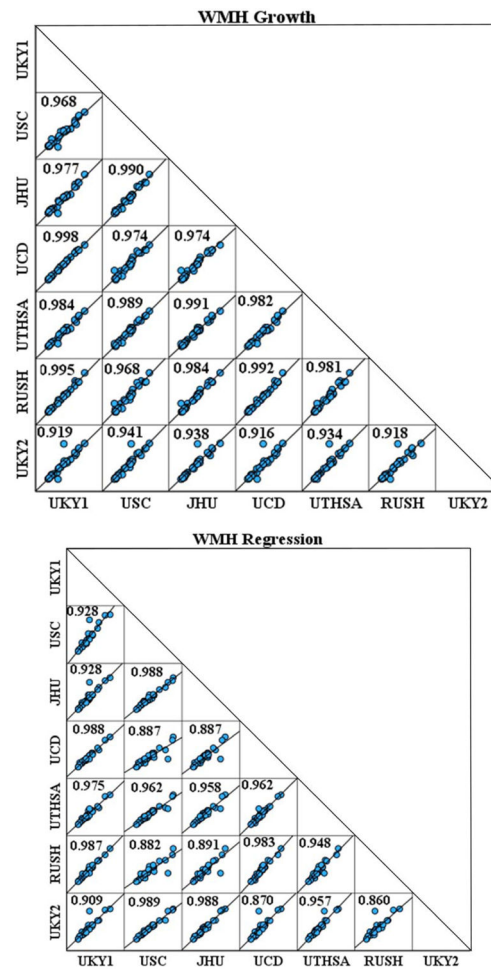
Fig. 3.
Intraclass correlation coefficient (ICC) matrix demonstrating robust correlations across raters both at the University of Kentucky (two raters) and across 5 other MarkVCID consortium sites for WMH A) Growth and B) Regression. UKY1, UKY2, University of Kentucky, raters 1 and 2; RUSH, Rush University Medical Center; JHU, Johns Hopkins University School of Medicine; USC, University of South California; UTHSCSA, University of Texas Health San Antonio; UCD, University of California, Davis.
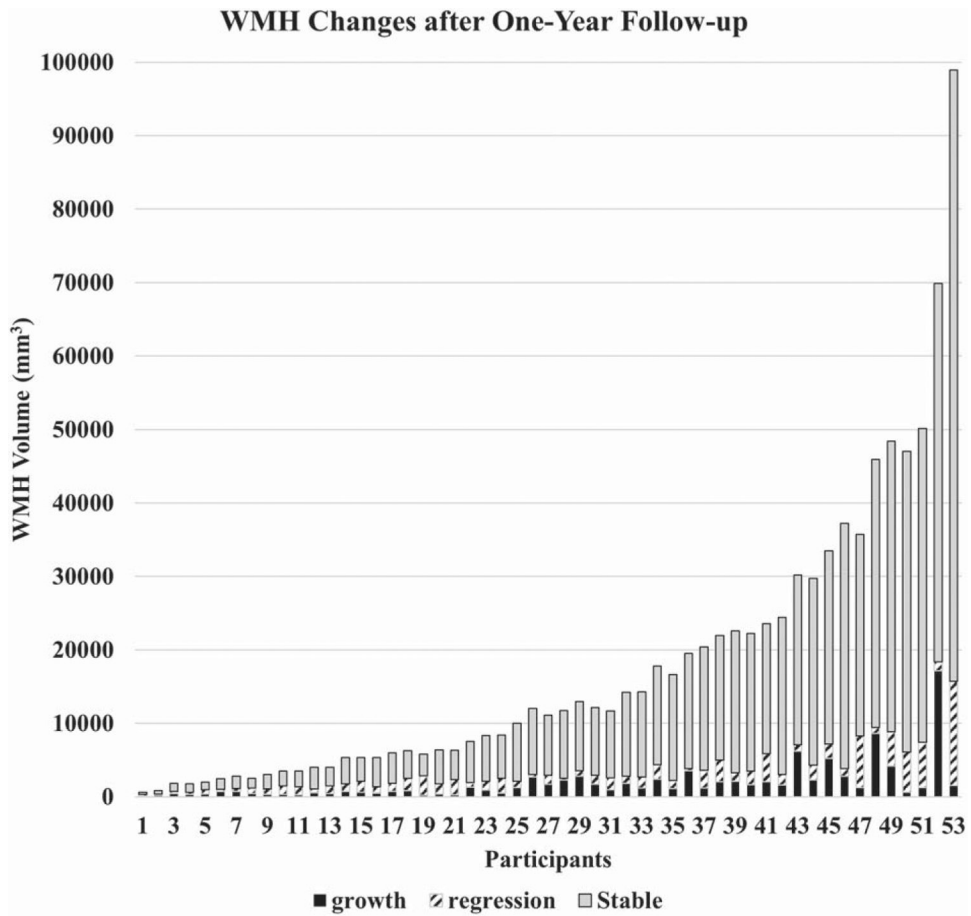
Fig. 4.
Histogram depiction of within-subject WMH change for the 53 participants that were analyzed for the preliminary assessment of clinical meaningfulness. WMH volume stability is dominant, while WMH volume growth and regression are variable across subjects and ranges of WMH volumes.
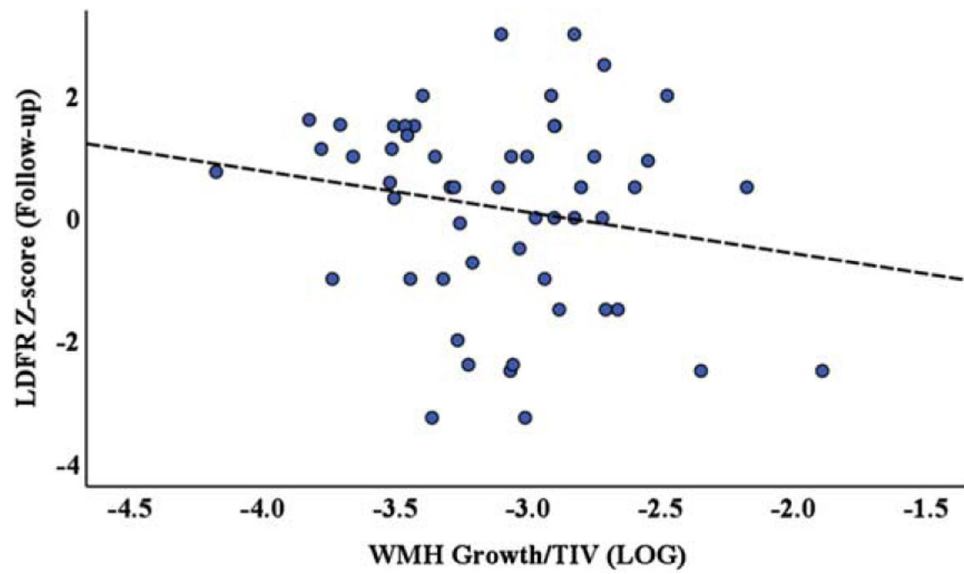
Fig. 5.
Scatter plot of log WMH volume growth normalized to the total intercranial volume versus LDFR z-score one-year follow-up scores, analyzed using a linear regression model adjusted for baseline LDFR z-score. The results demonstrate that WMH volume growth is significantly, negatively associated with CVLT-LDFR change ($p = 0.028$).

Table 1

Demographic and clinical variables for the participants analyzed for cross-site inter-rater and test-retest reliability/reproducibility assessments

Demographic and clinical variables	Cross-site inter-rater reliability (*n* = 30)	Test-retest repeatability (*n* = 18)	Preliminary analysis of construct validity (*N* = 53)
Age (y, mean ± SD)	75.8 ± 8.1	73.1 ± 1	75 ± 3
Education (y, mean ± SD)	19.7 ± 13.1	17 ± 2	16.4 ± 3.4
Sex (M:F)	17:13	2:1	29:24
Baseline MoCA Score (mean ± SD)	25.8 ± 2.6	25.7 ± 3.4	26 ± 2.7
Total WMH volume (cm³); baseline (mean ± SE)	11.10 ± 6.19	11.43 ± 0.24	15.7 ± 2.44
Total WMH volume (cm³): follow-up	10.85 ± 6.27	11.83 ± 0.19	15.76 ± 2.40
WMH growth volume (cm³)	1.41 ± 0.88	1.43 ± 0.08	1.84 ± 0.37
WMH regression volume (cm³)	1.65 ± 0.93	0.94 ± 0.06	1.79 ± 0.31
WMH Fazekas Score (Follow-up)	2.6 ± 0.9	2.06 ± 1.10	2.21 ± 0.89

SD, standard deviation; SE, Standard error; MoCA, Montreal Cognitive Assessment; WMH, white matter hyperintensities.