Validation and comparison of three freely available methods for extracting white matter hyperintensities: FreeSurfer, UBO Detector and BIANCA

Authors

Isabel Hotz [a,b], Pascal F. Deschwanden [a], Franziskus Liem [b], Susan Mérillat [b], Spyridon Kollias [d], Lutz Jäncke [a,b,c]

Author Affiliations

[a] Division of Neuropsychology, Department of Psychology, University of Zurich, Switzerland
[b] University Research Priority Program (URPP), Dynamics of Healthy Aging, University of Zurich, Zurich, Switzerland
[c] Department of Special Education, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
[d] Department of Neuroradiology, University Hospital Zurich, Zurich, Switzerland

Data and code availability statement

The data supporting this manuscript are not publicly available because the used consent does not allow for the public sharing of the data. We are currently working on a solution for this matter. Our code for preparing the data for BIANCA and executing BIANCA can be found here: https://github.com/dynage/bianca

The following openly available software were used:

FreeSurfer (https://surfer.nmr.mgh.harvard.edu/fswiki/DownloadAndInstall)

UBO Detector (https://cheba.unsw.edu.au/research-groups/neuroimaging/pipeline)

BIANCA (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Userguide), part of FSL software (RRID: SCR_002823, https://fsl.fmrib.ox.ac.uk) and the MATLAB implementation of LOCATE (https://git.fmrib.ox.ac.uk/vaanathi/LOCATE-BIANCA)

Funding statement

This research was funded by the Velux Stiftung (project No. 369) and the University Research Priority Program «Dynamics of Healthy Aging» of the University of Zurich (UZH).

Conflict of interest disclosure

The authors declare no conflicts of interest.

Ethics approval statement

The study was approved by the ethical committee of the canton of Zurich.

Patient consent statement

Participation was voluntary and all participants gave written informed consent in accordance with the declaration of Helsinki.

ORCID

Isabel Hotz: http://orcid.org/0000-0002-7938-0688

Corresponding Authors

Isabel Hotz
Binzmuehlestrasse 14, Box 25
CH-8050 Zurich
Switzerland
isabel.hotz@uzh.ch

Lutz Jäncke
Binzmuehlestrasse 14, Box 25
CH-8050 Zurich
Switzerland
lutz.jaencke@uzh.ch

Abstract

White matter hyperintensities of presumed vascular origin (WMH) are frequently found in MRIs of patients with neurological and vascular disorders, but also in healthy elderly subjects. Here, we compared and validated three freely available algorithms for WMH extraction: FreeSurfer, UBO Detector, and the Brain Intensity AbNormality Classification Algorithm (BIANCA).

The algorithms were applied to longitudinal MRI data of cognitively healthy older adults (baseline *N* = 231, age range 64 – 87 years). We fully manually segmented WMH in T1w, 3D FLAIR, 2D FLAIR MR images of 16 subjects and compared the corresponding algorithm outputs to assess the algorithm's segmentation accuracy. To validate the algorithms, we associated the extracted WMH volumes with the Fazekas scores and age, and explored correlations between repeated measures using the full longitudinal dataset. We further performed post hoc analyses of the algorithm's outputs to verify longitudinal robustness.

FreeSurfer fundamentally underestimated the WMH volumes and scored the worst in Dice Similarity Coefficient. Nevertheless, FreeSurfer's WMH volumes strongly correlated with the Fazekas scores and exhibited longitudinal robustness. BIANCA accomplished a high segmentation accuracy, but associations with the Fazekas scores were weak and a significant number of outlier-WMH volumes was identified in the within-person trajectories. UBO Detector achieved the best results in terms of the cost-benefit ratio. Although there is room for optimization, its performance was of high accuracy and consistency, despite being based on a built-in training dataset. In summary, our results emphasize the great importance of carefully contemplating the choice of the WMH segmentation algorithm in future studies.

Keywords

White matter hyperintensities
Automated segmentation
MRI
Healthy aging
Validation

1. Introduction

As our lifespan increases and the population ages, cognitive limitations caused by cerebrovascular diseases will become more common (Baker et al., 2012). White matter hyperintensities of presumed vascular origin (WMH) are considered as a marker of cerebrovascular diseases. They appear hyperintense on T2-weighted MRI images, like fluid-attenuated inversion recovery (FLAIR) sequences, without cavitation, and isointense or hypointense on T1-weighted (T1w) sequences (Wardlaw et al., 2013). The FLAIR sequence is generally the most sensitive structural sequence for visualizing WMH via magnetic resonance imaging (MRI) (Wardlaw et al., 2013). WMH are often seen in MR images of the brain in patients with neurological and vascular disorders but also in healthy elderly people (Caligiuri et al., 2015). In the context of clinical diagnostics, the Fazekas scale (Fazekas, Chawluk, Alavi, Hurtig, & Zimmerman, 1987), the Scheltens scale (Scheltens et al., 1993), and the age-related white matter changes scale (ARWMC) (Wahlund et al., 2001) are commonly used to visually assess the severity and progression of WMH. However, these visual rating scales unfortunately do not provide true quantitative data, have a relatively low reliability and are time-consuming to obtain (Mäntylä et al., 1997). Compared to such scales, volumetric measurements are more reliable and more sensitive to age effects in longitudinal studies of WMH (T. L. A. van den Heuvel et al., 2016), especially in cognitively healthy samples where the expected WMH volume increases over time are rather small. While several previous studies have been segmenting WMH manually (Anbeek, Vincken, van Osch, Bisschops, & van der Grond, 2004; Dadar et al., 2017; de Sitter et al., 2017; Klöppel et al., 2011; Kuijf et al., 2019; Steenwijk et al., 2013), this extremely time-consuming option seems not feasible in future studies, especially when considering the current trend towards big data (i.e., datasets with a large *N* and multiple time points of data acquisition). Automated methods that can detect WMH robustly and with high accuracy are therefore very useful and promising. Recently, Caligiuri and colleagues (2015) compared different existing algorithms including supervised, unsupervised, and semi-automated techniques. They found that many of these algorithms are not freely available, study and/or protocol specific, and have been validated on small sample sizes. There is still no consensus on which algorithm(s) is (are) of good quality and should be

applied to detect WMH (Dadar et al., 2017; Frey et al., 2019). Consequently, the methodology of pertinent studies is very heterogeneous and compromises the comparability of such studies.

The primary goal of our current work is therefore to validate and precisely compare the performance of three freely available WMH extraction methods: FreeSurfer (Fischl, 2012), UBO Detector (Jiang et al., 2018), and BIANCA (Brain Intensity AbNormality Classification Algorithm) (Griffanti et al., 2016). The FreeSurfer Image Analysis Suite (Fischl, 2012) is a fully automated software for the surface- and volume-based analysis of brain structure using information of T1w images. While quantifying WMH is not FreeSurfer's main aim, it still provides an unsupervised WMH segmentation as part of its pipeline and enables WMH quantification based on T1w images alone. UBO Detector and BIANCA are supervised tools specifically developed to automatically respectively semi-automatically segment WMH based on the k-nearest neighbor (KNN) algorithm. Although recent studies have been using these three algorithms, their validity and reliability is still not sufficiently clear. Former validations were often performed i) in patients with moderate to high WMH load, ii) using cross-sectional data, iii) in small samples. For a better overview see **Supplementary Table 1.**

In this study, we aim to provide complementary information on the performance of the three WMH extraction algorithms dependent of different MRI input modalities (T1w, 3D FLAIR + T1w, 2D FLAIR + T1w). To estimate the accuracy of WMH segmentation of the algorithms, we used fully manually segmented WMH as a proxy for true WMH. We refer to these manually segmented WMH as gold standards. The algorithms were applied to a longitudinal MRI dataset comprising cognitively healthy older adults (baseline $N = 231$, study interval = four years).

We addressed two study questions. The **first study question** evaluated and compared the WMH segmentation accuracy of the three algorithms using various accuracy metrics. To do so we manually segmented 16 subjects per modality. We refer to these manually segmented WMH as gold standards. The **second study question** examined the validity and reliability of the three algorithms by automatically extracting the WMH volumes from the entire dataset. As a clinical validation, the WMH volumes were associated with the Fazekas scale to explore if the algorithms are consistent with the visual assessment of WMH. The WMH volumes were further associated with chronological

age, as higher age has been consistently shown to be positively related to WMH load (D. M. J. van den Heuvel et al., 2006). Next, we explored correlations between the WMH volumes of consecutive time points to estimate the reliability of the algorithms.

2. Methods

2.1. Subjects

Longitudinal MRI data were taken from the Longitudinal Healthy Aging Brain Database Project (LHAB; Switzerland) – an ongoing project conducted at the University of Zurich (Zöllig et al., 2011). We used data from the first four measurement occasions (baseline, 1-year follow-up, 2-year follow-up, 4-year follow-up). The baseline LHAB dataset includes data from 232 participants (age at baseline: $M = 70.8$, range = 64 – 87; F:M = 114:118). At each measurement occasion, participants completed an extensive battery of neuropsychological and psychometric cognitive tests and underwent brain imaging. Inclusion criteria for study participation at baseline were age ≥ 64 years, right-handedness, fluent German language proficiency, a score of ≥ 26 points on the Mini Mental State Examination (Folstein, Folstein, & McHugh, 1975), no self-reported neurological disease of the central nervous system and no contraindications to MRI. The study was approved by the ethical committee of the canton of Zurich. Participation was voluntary and all participants gave written informed consent in accordance with the declaration of Helsinki.

For the present analysis, we only included participants with complete structural MRI data, which resulted in a baseline sample size of $N = 231$ (age at baseline: $M = 70.8$, range = 64 – 87; F:M = 113:118). At 4-year follow-up, the LHAB dataset still comprised 74.6% of the baseline sample ($N = 173$), of which $N = 166$ (age at baseline: $M = 74.2$, range = 68–87; F:M = 76:90) had complete structural data. In accordance with previous studies in this field, we ensured that none of the included datasets had intracranial hemorrhages, intracranial space occupying lesions, WMH mimics (e.g., multiple sclerosis), large chronic, subacute or acute infarcts, and extreme visually apparent movement artefacts.

2.2. MRI data acquisition

Longitudinally data MRI scans were acquired at the University Hospital of Zurich on a Philips Ingenia 3T scanner (Philips Medical Systems, Best, The Netherlands) using the dsHead 15-channel head coil. T1w and 2D-FLAIR structural images were part of the standard MRI battery, and are therefore available for the most time points (see **Table 1**). T1w images were recorded with a 3D T1w turbo field echo (TFE) sequence, repetition time (TR): 8.18 ms, echo time (TE): 3.799 ms, flip angle (FA): 8°, 160 × 240 × 240 mm^3 field of view (FOV), 160 sagittal slices, in-plain resolution: 256 × 256, voxel size: 1.0 × 0.94 × 0.94 mm^3, scan time: ~7:30 min. The 2D FLAIR image parameters were: TR: 11000 ms, TE: 125 ms, inversion time (TI): 2800 ms, 180 × 240 × 159 mm^3 FOV, 32 transverse slices, in-plain resolution: 560 × 560, voxel size: 0.43 × 0.43 × 5.00 mm^3, interslice gap: 1mm, scan time: ~5:08 min. 3D FLAIR images were recorded for a subsample only. The 3D FLAIR image parameters were: TR: 4800 ms, TE: 281 ms, TI: 1650 ms, 250 × 250 mm FOV, 256 transverse slices, in-plain resolution: 326 × 256, voxel size: 0.56 × 0.98 × 0.98 mm^3, scan time: ~4:33 min. **Table 2** provides an overview of the number of available MRI images per time point and per modality (T1w, 2D FLAIR, 3D FLAIR). While the 2D FLAIR + T1w images, and 3D FLAIR + Tw images serve as input for the validation of the UBO Detector and BIANCA algorithms, the T1w images only are used for the validation of the FreeSurfer algorithm.

Table 1

Number of subjects (N) broken down for modality (T1w, 2D FLAIR, 3D FLAIR) and time point (Tp).

	Time points				
Modality	**Tp1 *n***	**Tp2 *n***	**Tp3 *n***	**Tp5 *n***	**Total *N***
T1-weighted	231	207	196	166	800
2D FLAIR	228	203	174	157	762
3D FLAIR	4	46	53	63	166

2.3. Datasets used for algorithm input

In this work we used three datasets to validate the algorithms. The datasets differed in MRI sequence and number of subjects.

Table 2

Name of the datasets, input modality for the different algorithms, and the total number (N) of subjects per dataset.

Datasets	Algorithm(s)	Input modality(s)	Total *N* of subjects
Dataset 1	FreeSurfer	T1w	800
Dataset 2	UBO Detector & BIANCA	2D FLAIR + T1w	762
Dataset 3	UBO Detector & BIANCA	3D FLAIR + T1w	166

2.4. Accuracy metrics

We used various metrics to draw specific and comprehensive conclusions about the different segmentation accuracies of the algorithms. These metrics provide information about the degree of overlap, the degree of resemblance, and the volumetric agreement when comparing (a) the manually segmented reference – here named as gold standards – amongst each other and (b) the algorithm outputs with the gold standards. The equations can be found in **Table 3**. The **overlap agreement** is a commonly used metric representing the accuracy of classification. If voxels are correctly classified in a binary segmentation, they can be true positives (TPs) or true negatives (TNs). In contrast, when there is a discrepancy between the gold standard and the algorithm, then the voxels are false positives (FPs) or false negatives (FNs). In case of a low WMH load, as in this work, the number of TPs is much smaller than the number of TNs, which can affect the accuracy measures. The TPR, also known as sensitivity, was calculated and reported. The specificity (also known as recall) was not provided since it is equal to 1 - FPR. The **DSC** provides information about the overlap agreement between two segmentations (operators or automated segmentations) and it is perhaps the most established metric in evaluating the accuracy of WMH segmentation methods. However, since the DSC depends on the lesion load (the higher the lesion load, the higher the DSC), it is difficult to

evaluate operators or automated segmentation methods against each other if assessed on different sets of scans with different lesion loads (Wack et al., 2012). Extending the DSC, the **Outline Error Rate** (OER) and **Detection Error Rate** (DER) are independent of lesion burden (Wack et al., 2012). In these metrics, the sum of FP and FN voxels is split, depending on whether an intersection occurred or not. The sum is then divided by the mean total area (MTA) of the two operators to obtain a ratio with the DER as a metric of errors without intersection, and with the OER as a metric of errors with an intersection. We also applied the **Hausdorff distance** which is a shape comparison method, and can be used to evaluate the resemblance agreement of two images (Beauchemin, Thomson, & Edwards, 1998; Huttenlocher, Klanderman, & Rucklidge, 1993). It represents the maximum distance of a point in one set to the nearest point in the other set (Shonkwiler, 1991). To avoid problems with noisy segmentations we used the modified Hausdorff distance for the 95th percentile (H95) (Huttenlocher et al., 1993). We additionally calculated the volumetric agreement, which is a variant of the **Interclass Correlation Coefficient** (ICC). For comparisons with a gold standard we used the «unit» single (ICC(3,1)) and for the comparison without a gold standard, we used the equation with a pooled average (ICC(3,*k*)) (Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979).

Table 3

The following metrics were used to determine the agreements between the operators (Inter-Operator) and between the outcomes of the algorithms and the gold standards (Validation).

Metrics	Formulas (Inter-Operator)	Formulas (Validation)
Hausdorff distance for the 95th percentile (H95)	$d(A,B) = {}^{95}K^{th}_{a\epsilon A} d(a,B)$[a]	
Dice Similarity Coefficient (DSC)	$\frac{2 * V_{A \cap B}}{V_A + V_B}$	$\frac{2 * TP}{FP + 2 * TP + FN}$
Detection Error Rate (DER) (All Clusters without intersection $V_{A \cap B}$ divided by MTA[b])	$\frac{V_A + V_B}{MTA}$	$\frac{2 * (FP + FN)}{\mathrm{FP} + \mathrm{FN} + 2 * \mathrm{TP}}$
Outline Error Rate (OER) (All Clusters with intersection $V_{A \cap B}$ divided by MTA)	$\frac{V_A + V_B - 2 * V_{A \cap B}}{MTA}$	$\frac{2 * (FP + FN)}{\mathrm{FP} + \mathrm{FN} + 2 * \mathrm{TP}}$

Sensitivity = True positive ratio (TPR)	$\frac{TP}{TP + FN}$
False positive ratio (FPR)	$\frac{FP}{FP + TN}$

Notes: [a] $^{95}K^{th}_{a\epsilon A}$ represents the Kth ranked distance such that $K/N_a = 95\%$ (Dubuisson & Jain, 1994).
MTA = mean total area, area of rater A and area of rater B divided by 2 (Wack et al., 2012).

2.5. Manual WMH segmentation

To provide a valid baseline for the evaluation of the WMH segmentation accuracy of the different algorithms, WMH were manually segmented in a subsample of 16 subjects. We refer to these WMH segmentations as the gold standards given that the manually segmented WMH represent strong proxies for the true WMH load. Only participants, for which T1w, 3D FLAIR and 2D FLAIR images were available for all four time points, were considered for the manual segmentation. For each of the 16 subjects, 3 images - one of each modality (T1w, 3D FLAIR, 2D FLAIR) - were manually segmented, thus, resulting in a total of 48 binary gold standard masks with the values *0* for background and *1* for WMH. In order to represent the entire dataset as well as possible, and to provide a useful training dataset for BIANCA, individual Fazekas scores were considered to ensure that the subsample chosen for manual segmentation covered a mixed range of WMH loads but at the same time had a medium Fazekas score on average (see validation of the gold standard).

Segmentation of FLAIR images: Three operators (O1, O2, O3) completely manually segmented the WMH of 16 subjects with 3D FLAIR images and of ten subjects with 2D FLAIR images in all three planes (sagittal, coronal, axial) on a MacBook Pro 13-inch with a Retina Display with a screen resolution of 2560 × 1600 pixels, 227 pixels per inch with full brightness intensity to obtain comparable data using FSLeyes (McCarthy, 2018). The segmentations were carried out independently resulting in three different masks per segmented image. Because of the very good preceding segmentation-inter-operator reliabilities, only one operator (O2) segmented the remaining six subjects to achieve the same number of gold standards for all modalities. Nevertheless, to ensure high segmentation quality, these WMH masks were peer-reviewed by O1, O3 checked all images again, and any discrepancies were discussed between all operators and S.K. Before and during segmentation, all operators were trained and supported for several months by *S.K.*, a neuroradiology

professor with over 30 years of experience in diagnosing cerebral MR images. The three manually segmented WMH masks of the same subjects (FLAIR images) of the three operators were displayed as overlays in FSLeyes in order to evaluate a mask agreement across these three operators (voxel value *1.0*: all three operators classified the voxel as WMH; voxel value $0.\overline{666}$: two operators classified the voxel as WMH; $0.\overline{333}$: one operator classified the voxel as WMH (see **Supplementary Figure 1**). Each mask overlay was then revised voxel by voxel by consensus in the presence of all operators (O1, O2, O3), and converted back to a binary mask to serve as gold standard. The between-operator disagreements mostly regarded voxels at the WMH borders. The resulting masks were shown to *S.K.* and corrected in case of mistakes.

Segmentation of T1w images: Two operators (O1, O2) split the 16 subjects with the T1w images to carry out fully manual segmentation. All images were checked by O3, and any discrepancies were discussed amongst all operators and *S.K.*.

Validation of the gold standards: The mean DSC between all three operators for the 3D (n = 16) and 2D FLAIR images (n = 10 subjects) was 0.73 and 0.67, respectively. A mean DSC of 0.7 (Anbeek et al., 2004; Caligiuri et al., 2015) is considered as a good segmentation. As expected, due to the lower surface-to-volume ratio, the DSC was lower for images with a low WMH load than for images with a high WMH load (Wack et al., 2012). In images with a low WMH load, a DSC above 0.5 is still considered as a very good agreement (Dadar et al., 2017). Our average DSC results in both modalities (3D and 2D FLAIR images) were higher than 0.7 for medium WMH load, and higher than 0.6 for low WMH load, which can be considered as excellent agreement. The reliability of the volumetric agreement between the segmentations of the 3D and 2D FLAIR images, as indicated by the ICC, was excellent (Cicchetti, 1994) (3D FLAIR: mean ICC = 0.964; 2D FLAIR: mean ICC = 0.822). Detailed results on further metrics, and on segmented WMH volume can be found in **Supplementary Table 2**.

In preparation for the optimization phase of UBO Detector, and for the mandatory training dataset for BIANCA, six additional subjects with 2D FLAIR images (the remaining n = 6 subjects to obtain a subsample size of n = 16 subjects per modality) were manually segmented. Because of the very good

segmentation-inter-operator reliabilities in the first ten subjects with 2D FLAIR images, only one operator (O2) segmented the additional images. To assure high segmentation quality, the WMH masks for these six images were peer reviewed by O1. Then, O3 checked all images, and any discrepancies were discussed amongst all operators and *S.K.*. **Table 5** (first row) shows an overview of the mean WMH volumes resulting from the manual segmentations based on the 16 subjects per modality (T1w, 3D FLAIR, 2D FLAIR). No significant mean WMH volume differences were revealed between the three different modalities using the Kruskal-Wallis test ($X^2_{(2)} = 0.0016$, $p = 0.999$). Also, after the post-hoc test (Dunn-Bonferroni with Holm correction) there were no significant WMH volume differences between the manually segmented gold standards. The Pearson's product-moment correlation showed an almost perfect (Dancey & Reidy, 2017) linear association between all gold standards (all combinations): mean (0.97, $p < 0.001$) (see **Supplementary Figure 2** for the correlations between the gold standards but also between all three algorithms per input modality).

2.6. Fazekas scale

The Fazekas scale is a widely used visual rating scale that provides information about the location (periventricular WMH (PVWMH) and deep WMH (DWMH)) and the severity of WMH lesions. It ranges from 0 to 3 for both locations, leading to a possible minimum score of 0, and a maximum score of 6 for total WMH. In a first step, the three operators (O1, O2, O3) were specially trained by the neuroradiologist *S.K.* for several weeks on evaluating WMH with the Fazekas scale. *S.K.* was blinded to the demographics and neuropsychological data of the participants. 800 images were then visually rated using the Fazekas scale. For this purpose, we used the 3D and 2D FLAIR images. In rare cases, when no FLAIR images were available, the T1w were used as a basis. The ratings were carried out independently by all three operators, validated for the further procedure with statistical indicators that are described as follows. The inter-operator mean concordances across all four time points were determined with Kendall's coefficient of concordance (Moslem, Ghorbanzadeh, Blaschke, & Duleba, 2019) by calculating it for total WMH, DWMH and PVWMH separately for

each time point. Strong (Moslem et al., 2019) agreements for total WMH ($W = 0.864$, $p < 0.001$), PVWMH ($W = 0.828$, $p < 0.001$), and DWMH ($W = 0.842$, $p < 0.001$) were found. Furthermore, mean inter-operator reliabilities were evaluated between the three operators for total WMH, PVWMH and DWMH across the four time points by using a weighted Cohen's kappa (Cohen, 1968). A substantial to near-perfect (Landis & Koch, 1977) reliability was found (see **Supplementary Table 3**). Finally, the Spearman's rho was calculated to investigate the rank correlation of the extracted WMH volumes of the different algorithms and the Fazekas scores. This choice was made based on to the skewed distribution of the WMH volumes, and the fact that the Fazekas scale is ordinally scaled. Therefore, the median score was calculated for each participant for each time point, split into total WMH, PVWMH and DWMH. The median Fazekas scores for all three operators were: total WMH = 3; PVWMH = 2; DWMH = 1. For more descriptive details see (**Supplementary Table 4**).

2.7. Automated WMH segmentation

In order to obtain the best performance from each algorithm, we have chosen the settings that were considered the best by the original authors from UBO Detector and BIANCA to lead to the best results. Thus, it is not a question of examining comparable settings (e.g., for the PVWMH), but rather the outcome of the algorithms under optimal conditions.

2.7.1. FreeSurfer

The FreeSurfer Image Analysis Suite (Fischl, 2012) uses a structural segmentation to identify regions in which WMH can occur, while regions in which WMH cannot occur are excluded (cortical and subcortical gray matter structures). The algorithm assigns a label to each voxel based on probabilistic local and intensity-related information that is automatically estimated from a built-in training dataset comprising 41 manually segmented images (Fischl et al., 2002). The algorithm differentiates between hypointensities in the white (WMH) and grey matter (non-WMH). The T1w images (dataset 1) were processed with FreeSurfer v6.0.1 as implemented in the FreeSurfer BIDS-App (Gorgolewski et al., 2017).

2.7.2. UBO Detector

The UBO Detector WMH extraction pipeline uses T1w and FLAIR images as input. We conducted the analysis once with dataset 2 (2D FLAIR + T1w), and once with dataset 3 (3D FLAIR + T1w). The probability of WMH is calculated by applying a classification model trained on ten subjects, manually segmented on 2D FLAIR images (built-in training dataset). A user-definable probability threshold generates a WMH map by segmenting the subregions including PVWMH, DWMH, lobar and arterial regions. As recommended by Jiang and colleagues (2018) we used a 12 mm threshold to define the borders of PVWMH. The segmentation of the T1w images failed in five images, whereupon these subjects were excluded from the following procedures. Visual inspection of the data uncovered a segmentation error (eyeballs have been marked as WMH), thus, this time point was excluded from further analysis leading to a total number of $N = 756$ subjects. To determine which settings are best suited for our datasets, we have evaluated the performance of four different settings proposed by Jiang et al. (2018), using the leave-one out cross-validation method. Since we had manually segmented the WMH for both, 3D and 2D FLAIR images, we examined the performance of UBO Detector separately for each modality (T1w, 3D FLAIR + T1w, 2D FLAIR + T1w). To do so, we calculated the accuracy metrics separately for the different settings, and checked which adjustments achieved the most optimal values (for further results see **Supplementary Table 5**). For 3D FLAIR images, UBO Detector worked most accurately with a threshold of 0.7 and a NN of $k = 5$. For the 2D FLAIR images, the best performance was achieved with a threshold of 0.9 and a NN of $k = 3$. For the subsequent calculations we used these optimized settings.

2.7.3. BIANCA

For BIANCA (Griffanti et al., 2016) a training dataset is mandatory. As an output, BIANCA generates a probabilistic map of WMH for total WMH, PVWMH and DWMH. We performed the calculations once with dataset 2 (2D FALIR + T1w) and once with dataset 3 (3D FLAIR + T1w) – using the leave-one out cross-validation method. The 16 manually segmented gold standards derived once from 3D and once 2D FLAIR images were used for the training datasets combined with the T1w images. For defining the PVWMH we adopted the 10 mm distance rule from the ventricles

380 (DeCarli et al., 2005), which was also suggested by Griffanti et al. (2016). To reduce false positive
381 voxels in the grey matter, and at the same time only localize WMH in the white matter, we applied a
382 WM mask. For the BIANCA options we chose the ones that Griffanti et al. (2016) indicated as the
383 best in terms of DSC and cluster-level false-positive ratio: MRI modality = FLAIR + T1w, spatial
384 weighting = «1», patch = «no patch», location of training points = «noborder», number of training
385 points = number of training points for WMH = 2000 and for non-WMH = 10000. For more details on
386 the descriptions and the options, see Griffanti and colleagues (2016).
387 The preprocessing steps, applied before the BIANCA segmentation procedure, were performed with
388 a nipype pipeline (v1.4.2; (Gorgolewski et al., 2011) as follows: Based on subject-specific template
389 created by the anatomical workflows of fMRIprep (v1.0.5; (Esteban et al., 2019)), a WM-mask
390 (FSL's make_bianca_mask command) and a distancemap (distancemap command) were created. The
391 distancemap was thresholded into periventricular and deep WM (cut-off = 10 mm). For each session,
392 T1w images were bias-corrected (ANTs v2.1.0; (Tustison et al., 2010)), brought to the template
393 space, and averaged. FLAIR images were bias-corrected, and the template-space images were
394 brought into FLAIR space using FLIRT (Jenkinson & Smith, 2001). To select the best threshold for
395 the probabilistic output of BIANCA we first used the leave-one out cross-validation method to
396 calculate the different validation metrics separately for the 3D and 2D FLAIR gold standard images
397 (+ T1w images). The global threshold values 0.90, 0.95, 0.99 were applied, with the threshold of 0.99
398 for both FLAIR sequences proving to be the best fitting. For a more detailed overview see
399 **Supplementary Table 6**.

400 2.8. Analysis plan

401 In preparation of the main analyses, we carried out two initial analysis steps. **First**, we evaluated the
402 segmentation's accuracy of the 16 gold standards per modality (T1w, 2D FLAIR, 3D FLAIR) by
403 using the following accuracy metrics: DSC, DER, OER, H95, and ICC. The DCS for the FLAIR
404 images was calculated for different levels of WMH loads (for the results see **Supplementary Table**
405 **2**). **Secondly**, we compared the two thresholding options of BIANCA – best global threshold (0.99 –

evaluated before, see **2.7.3 BIANCA**) vs. LOCally Adaptive Threshold Estimation (LOCATE) to investigate which option provides the better segmentation quality for our dataset. To do so we used the leave-one out cross-validation method with the 16 gold standards per modality. We compared the outputted WMH with the DSC, DER, OER, H95, FP, TP, FPR, Sensitivity, ICC but also with the WMH volumes in cm^3 (see **Supplementary Table 10)** by using the Wilcoxon rank sum test. Because LOCATE did not perform better than the best global thresholding with 0.99 we used the latter for further analyses. For more information, and the detailed comparison analysis, see **Supplementary Analysis.**

The main analysis comprised two parts in line with our two research questions. **For the first study question**, the accuracies of the three algorithms (FreeSurfer, UBO Detector, and BIANCA) were calculated by comparing their outputs with the gold standards of the corresponding modality (T1w, 3D FLAIR, 2D FLAIR) using the accuracy metrics DSC, OER, DER, TPR, Sensitivity, ICC but also the WMH volumes. The comparisons with the DSC, H95, DER, OER, FPR, and Sensitivity were done using the Kruskal-Wallis, and analyzed post-hoc with the Dunn-Bonferroni (Holm correction). The ICC comparisons were interpreted with a degree of reliability according to Cicchetti (1994). The comparisons between the mean WMH volumes of the gold standards with the different modalities, as the comparisons between the outputted WMH volumes of the algorithms of the same subjects, also per modality, were performed with the Kruskal-Wallis test (post-hoc Dunn-Bonferroni with Holm correction). The comparisons between the mean WMH volumes of the 16 gold standards per modality and the mean WMH volumes of the same 16 subjects outputted by the algorithms per modality were calculated using the Wilcoxon-rank-sum-test and interpreted with a Cohen's *d* (Cohen, 1988). **For the second study question**, we investigated the associations of the log-transformed and z-standardized WMH volume outputs with (a) z-standardized Fazekas scores for clinical validation, and (b) z-standardized chronological age using linear mixed models (LMMs). Data were nested across time points. We used the Akaike Information Criterion (AIC) for the selection of the appropriate statistical models for each algorithm. Lower AIC values indicating better fitting quality; therefore, we chose the model that has the smallest AIC value. The random intercept model (model

2) revealed the lowest AIC for all calculations. Relative effect sizes were calculated following the tutorial from Brysbaert and Stevens (2018) and according to Westfall, Kenny, and Judd (2014). The analyses were carried out in R (R Core Team, 2020) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). To calculate the degrees of freedom, a Satterthwaite fit was used for all models. Since we were interested in the linear associations, we reported the fixed effects ($\hat{\beta}$) of the WMH volumes of the different algorithms, and the respective variables (i.e., Faszekas scores and chronological age). In addition, we explored correlations between the WMH volumes of consecutive time points. For the validation of FreeSurfer, dataset 1 (T1w) was used, while the validation of UBO Detector and BIANCA relied on dataset 2 (2D FLAIR + T1w) and 3 (3D FLAIR + T1w).

Finally, we ran **post-hoc calculations** based on our observations of strong discrepancies between WMH volumes of consecutive time points in the BIANCA output, which we refer to as «suspicious intervals» in the following. To further investigate the within-subject variability of WMH volumes, we determined the percentage and number of «suspicious intervals between two measurement points» and also of «subjects with suspicious longitudinal data» based on the mean percentages of WMH volume increases (mean + 1SD) and decreases (mean - 1SD), separately for possible time intervals (1-, 2-, 3-, 4-year intervals) to calculate a «range of tolerance» and exclude suspicious data points. The suspicious data points were further classified for low, medium, and high WMH load (based on the Fazekas scale) in order to identify a specific pattern. If the number of «suspicious intervals between two measurement points» exceeded the number of «subjects with suspicious longitudinal data», this indicated peaks or even several suspicious data points in a single person – and would be an indication of the zigzag pattern over time. For a more detailed description see **3.3. Longitudinal post hoc analyses of the algorithms' output.**

2.9. Computer equipment

All WMH extractions were undertaken on a Supermicro X8QB6 workstation with 4 × Intel Xeon E57-4860 CPU (4 x10 cores, 2.27 GHz) and 256 GB RAM. The computing host was a KVM

virtualized guest instance with Ubuntu 18.04.4 LTS with 32 x Intel Xeon E7-4860 CPU (2.27 GHz) and 92 GB RAM.

3. Results

The results are divided into two subsections. First, we report the performance of the algorithms compared to the 16 manually segmented gold standards per modality (T1w, 3D FLAIR, 2D FLAIR) using the different accuracy metrics. Further, we present the WMH volumes of the gold standards and those of the algorithms and their comparisons with each other in a table. Secondly, we represent the validation of the three algorithms with the entire longitudinal datasets (dataset 1 for FreeSurfer, dataset 2 and 3 for UBO Detector and BIANCA).

3.1. Segmentation performance of the algorithms

As demonstrated in **Table 4**, the DSC calculated from all algorithm's outputs is sensitive to the WMH load, categorized according to the respective manually segmented gold standard per modality (T1w, 3D FLAIR, 2D FLAIR). However, because of the constant underestimation of the WMH volume by FreeSurfer, it already reached the maximum DSC at the medium WMH load. Therefore, FreeSurfer could not achieve an improvement of the DSC from medium to high WMH exposure in terms of DSC.

Table 4

Comparison of the mean Dice Similarity Coefficient (DSC) of the three algorithms (FreeSurfer T1w, UBO Detector and BIANCA 3D FLAIR + T1w and 2D FLAIR + T1w) according to subjects with low (L) (< 5cm^3), medium (M) (5 – 15 cm^3), and high (H) (> 15 cm^3) mean WMH load for the corresponding manual segmentation of the same 16 subjects per modality (T1w, 3D FLAIR, 2D FLAIR).

WMH load		T1-w[a]	3D FLAIR[b]		2D FLAIR[c]	
		FreeSurfer	UBO Detector	BIANCA	UBO Detector	BIANCA
		Mean DSC				
< 5 cm^3	**L**	0.371	0.414	0.501	0.432	0.482
5 – 15 cm^3	**M**	0.487	0.550	0.674	0.556	0.577
> 15 cm^3	**H**	0.473	0.638	0.700	0.668	0.684

Total	0.434	0.501	0.602	0.531	0.561
± SD	± 0.111	± 0.124	± 0.133	± 0.113	± 0.118

Notes: [a] T1-w gold standard: L: (*n* = 7), M: (n = 7), H: (*n* = 2). [b] 3D FLAIR gold standard: L: (*n* = 7), M: (*n* = 7), H: (*n* = 2).[c] 2D FLAIR gold standard: L: (*n* = 5), M: (n = 9), H: (*n* = 2).

Table 5 shows the results of the different WMH volume comparisons of the gold standards and the algorithms. No significant differences were found between the mean WMH volumes of the three modalities T1w, 3D FLAIR and 2D FLAIR of the manually segmented gold standards. However, looking at the WMH volumes of the algorithm's outputs of the three input modalities, the WMH volumes of FreeSurfer are significantly smaller than those of all the other algorithms (see comparison between algorithms in **Table 5**). Comparing the WMH volumes of the gold standards with those from the automated algorithms, we found that FreeSurfer underestimated the same 16 images with a large effect size ($W = 43$, $p < 0.001$). FreeSurfer was thus the only algorithm to underestimate its «own» gold standard for the same subjects and MR images.

Table 5

Mean WMH volume in cm^3 of the manually segmented gold standards (N = 16 per image modality) and corresponding mean WMH volumes automatically estimated by the three algorithms. WMH volume comparisons were carried out amongst the gold standards, amongst the algorithm outcomes, and between the gold standards and the algorithm outcomes. Gray shaded cells indicate significantly smaller WMH volume or large effect size, bold cells indicate significantly larger WMH volume in comparison.

	T1w	**3D FLAIR + T1w**	**2D FLAIR + T1w**	**Comparisons between gold standards**	**Comparisons between algorithms**
	Mean volume in cm^3 (SD)	**Mean volume in cm^3 (SD)**	**Mean volume in cm^3 (SD)**	Kruskal-Wallis, post-hoc, Dunn-Bonferroni (Holm correction)	
Gold standards	7.781 (5.224)	8.311 (5.811)	9.032 (7.723)	$X^2_{(2)} = 0.0016$, $p = 0.999$	
FreeSurfer	3.436 (2.556)				FreeSurfer < UBO 2D $p = 0.006$
UBO Detector		6.001 (5.124)	**9.510 (6.704)**		FreeSurfer < BIANCA 3D $p = 0.028$
BIANCA		**8.317 (5.490)**	**8.695 (7.003)**		FreeSurfer < BIANCA 2D $p = 0.040$
Volume GS vs. Volume algorithm	**Wilcoxon, post-hoc Dunn-Bonferroni (Holm correction)**			**Interpretation with effect sizes according to Cohen's *d*†**	
	$W = 43$ $p < 0.001$	**UBO 3D:** $W = 85$ $p = 0.11$ **BIANCA 3D:** $W = 133$ $p = 0.743$	**UBO 2D:** $W = 137$ $p = 0.752$ **BIANCA 2D:** $W = 123$ $p = 0.827$	large effect size (d > 1.0)† between the mean volume of the T1w gold standard and the mean volume of FreeSurfer's output.	

Notes: GS = gold standard. Between the gold standards with the different modalities T1w, 3D FLAIR and 2D FLAIR no significant differences were found (Kruskal-Wallis, post-hoc Dunn-Bonferroni with

Holm correction). Between the algorithms (with the different modalities) FreeSurfer's volume was significant smaller than UBO Detector 2D FLAIR, BIANCA 2D FLAIR and BIANCA 3D FLAIR (Kruskal-Wallis, post-hoc Dunn-Bonferroni with Holm correction).
† Cohen's $d = \geq 0.2 - \leq 0.5$ = small = $0.5 - < 0.8$ = moderate, $0.8 - \geq 1.0$ = large effect size (Cohen, 1988) ; > 2.0 = huge effect size (Sawilowsky, 2009).

Table 6 provides an overview of the outcomes of the inferential statistics: The results of the different accuracy metrics for the three algorithms are compared, using the manually segmented gold standards for each modality. The results of the mean DSCs suggest that BIANCA performed best in terms of the degree of overlap between gold standards and algorithm output. However, a significant difference could only be seen between FreeSurfer and BIANCA. The 3D FLAIR inputs indicate a slightly better DSC value on average, with BIANCA 3D FLAIR showing the best result. Nevertheless, the algorithms using FLAIR + T1w images as input modality are comparable regarding the mean DSC. With the H95, a comparison is only valid within the same modality. Here BIANCA, especially BIANCA with the 3D FLAIR + T1w input, was very accurate, and on average, significantly better than UBO 3D FLAIR. Visually, there are also noticeable differences to the other algorithms, see **Supplementary Figure 3**. The accuracy of BIANCA was also demonstrated by the higher sensitivity of the 3D FLAIR + T1w images modality input compared to the 3D FLAIR + T1w images modality input of UBO Detector. However, UBO 2D FLAIR + T1w images does not differ significantly from both BIANCA modality inputs. Only FreeSurfer performed significantly worse than BIANCA and UBO 2D FLAIR. In contrast, FreeSurfer showed on average a significantly better FPR than all others, while it scored significantly worse in terms of OER compared to all others. FreeSurfer was the only algorithm that on average significantly underestimated the WMH volumes compared to the gold standard. Within BIANCA, the input modality 2D FLAIR + T1w performed worse than the 3D FLAIR + T1w input with a moderate effect size. A large effect size was shown between BIANCA 2D FLAIR + T1w images and FreeSurfer. UBO 2D FLAIR + T1w and UBO 3D FLAIR + T1w showed the highest correlation of all algorithms between the total WMH volume and the Fazekas scores.

Table 6

Summary of statistical comparisons (n = 16 per modality) of the accuracy metrics of the three algorithms per modality (T1w, 3D FLAIR, 2D FLAIR) using the Dice Similarity Coefficient (DSC), Hausdorff distance for the 95th percentile (H95), sensitivity, false positive ratio (FPR), Detection Error Rate (DER), Outline Error Rate (OER), and Interclass Correlation Coefficient (ICC). Gray shadowed cells indicate significantly worse performance or a worse degree of reliability.

	FreeSurfer (T1w)		**UBO (3D FLAIR)**		**UBO (2D FLAIR)**		**BIANCA (3D FLAIR)**		**BIANCA (2D FLAIR)**			
	Mean	**95% CI**	**Mean**	**95% CI**	**Mean**	**95% CI**	**Mean**	**95% CI**	**Mean**	**95% CI**	***p*-value**	**Post-hoc Dunn-Bonferroni (Holm correction)**
DSC[a]	0.434	0.375 – 0.493	0.500	0.435 – 0.567	0.531	0.471 – 0.591	0.602	0.531 – 0.672	0.561	0.498 – 0.624	0.005	BIANCA 3D > FreeSurfer
H95 (mm)[a]	8.660	6.804 – 10.515	11.533	8.520 – 14.545	13.522	10.324 – 16.720	6.200	4.093 – 8.305	8.443	6.421 – 10.464	< 0.001	BIANCA 3D > UBO 3D, UBO 2D
Sensitivity[a]	0.315	0.260 – 0.370	0.427	0.353 – 0.501	0.572	0.500 – 0.643	0.611	0.538 – 0.683	0.575	0.492 – 0.657	< 0.001	UBO 2D, BIANCA 3D, BIANCA 2D > FreeSurfer; BIANCA 3D > UBO 3D
FPR[a]	4.62 E^{-5}	0.0000 – 0.0001	0.0002	0.0001 – 0.0002	0.0004	0.0003 – 0.0006	0.0003	0.0001 – 0.0004	0.0004	0.0002 – 0.0005	< 0.001	FreeSurfer < all others; UBO 3D < UBO 2D
DER[a]	0.200	0.146 – 0.253	0.313	0.228 – 0.398	0.327	0.231 – 0.423	0.162	0.113 – 0.210	0.215	0.143 – 0.287	< 0.001	BIANCA 3D < UBO 3D, UBO 2D
OER[a]	0.932	0.839 – 1.025	0.687	0.629 – 0.746	0.611	0.540 – 0.683	0.635	0.538 – 0.733	0.664	0.586 – 0.742	< 0.001	all others < FreeSurfer
	coeff	***p*-value**	**coeff**	***p*-value**	**coeff**	***p*-value**	**coeff**	***p*-value**	**coeff**	***p*-value**	**Interpretation**	
ICC(3,1)	0.454	0.081	0.876	< 0.001	0.927	< 0.001	0.743	< 0.05	0.859	< 0.001	FreeSurfer fair, BIANCA 3D good, BIANCA 2D, UBO 3D and UBO 2D excellent degree of reliability†	

Notes: [a] Comparison by Kruskal-Wallis. [b] Comparison by Wilcoxon-rank-sum-test.
†Between 0.40 – 059 = fair, 0.60 – 0.74 = good, 0.75 and 1.00 = excellent (Cicchetti, 1994)

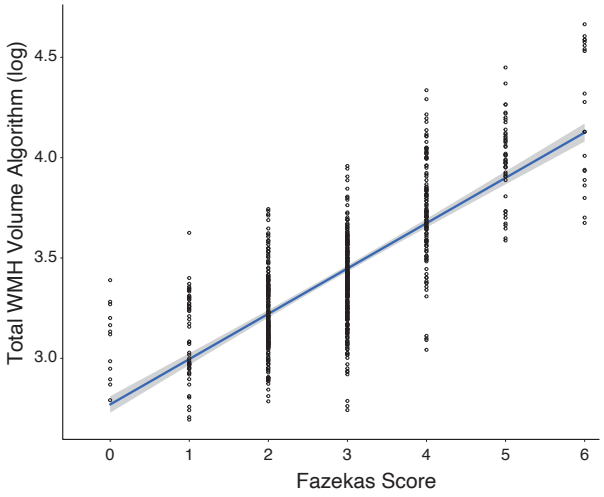
3.2. Validation of the algorithms

To validate the algorithms, we used LMMs to calculate the associations between the log-transformed and z-standardized WMH volumes with the z-standardized **Fazekas scores**, but also with z-standardized **chronological age**. As mentioned before (see **2.8 Analysis plan, second study question)**, the random intercept model (model 2) performed best in all calculations**.** To analyze the reliability of the WMH volumes over time, we correlated the WMH volumes **between the time points** (Tp1–Tp2, Tp2–Tp3, Tp3–Tp5). Figures 1, 2 and 3 illustrate the different validation steps for FreeSurfer, UBO Detector and BIANCA, respectively. The scatter plots illustrate the distribution of the total WMH volumes and the Fazekas scale of the respective datasets. The spaghetti plots show the WMH volume trajectories within each subject over time (chronological age). The plots depicting correlations of WMH volumes between two time points show how reliably the WMH volumes are estimated across time points. The corresponding results of all algorithms are shown on **Table 7.**

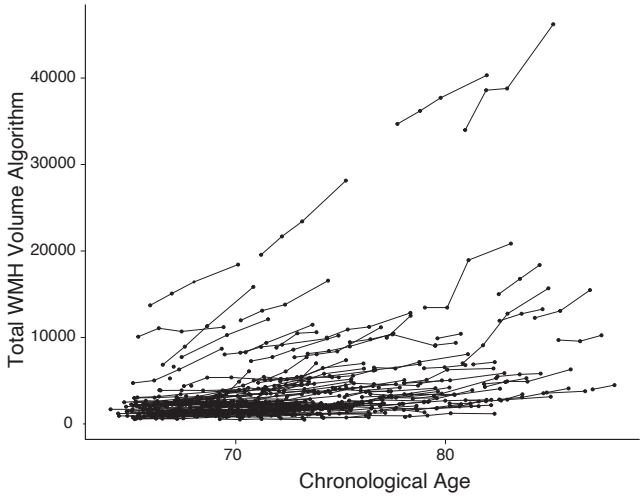
564 **Figure 1**

565 ***FreeSurfer validation.*** *Scatter plot of the log-transformed total WMH volume distribution according*
566 *to Fazekas scale with dataset 1 (A), spaghetti plot of total WMH volume and chronological age of*
567 *dataset 1(B), and the Pearson's product-moment correlation between the time points with dataset*
568 *1(C).*

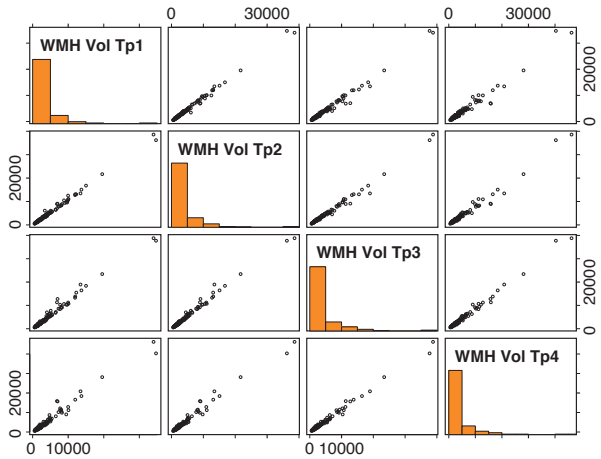
A) Dataset 1 – Total WMH Volume vs Fazekas Scale



B) Dataset 1 – Total WMH Volume vs Chronological Age



C) Dataset 1 – Correlation between Time Points



569

570

571

572

573

574

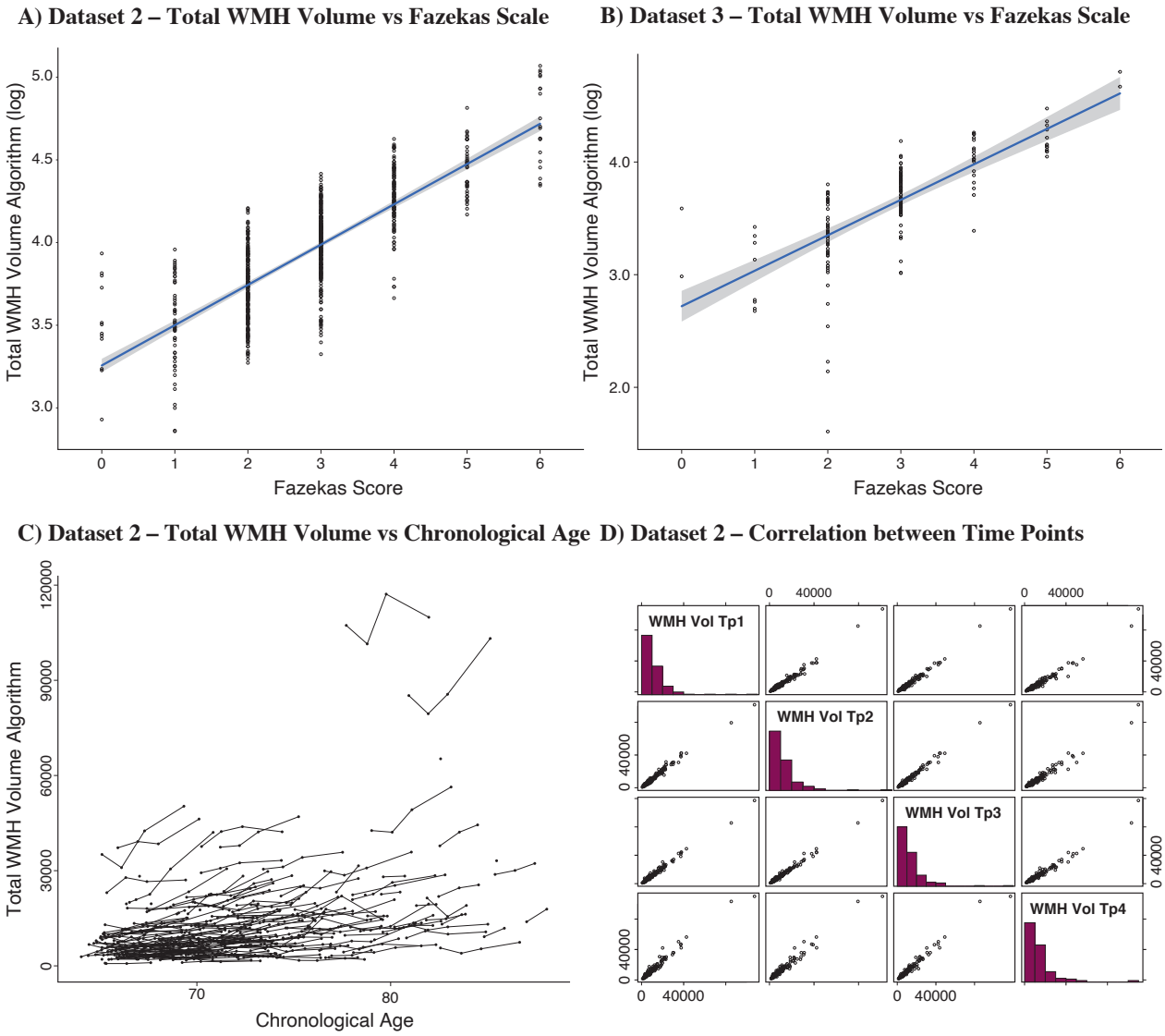
575

576

577

578 **Figure 2**

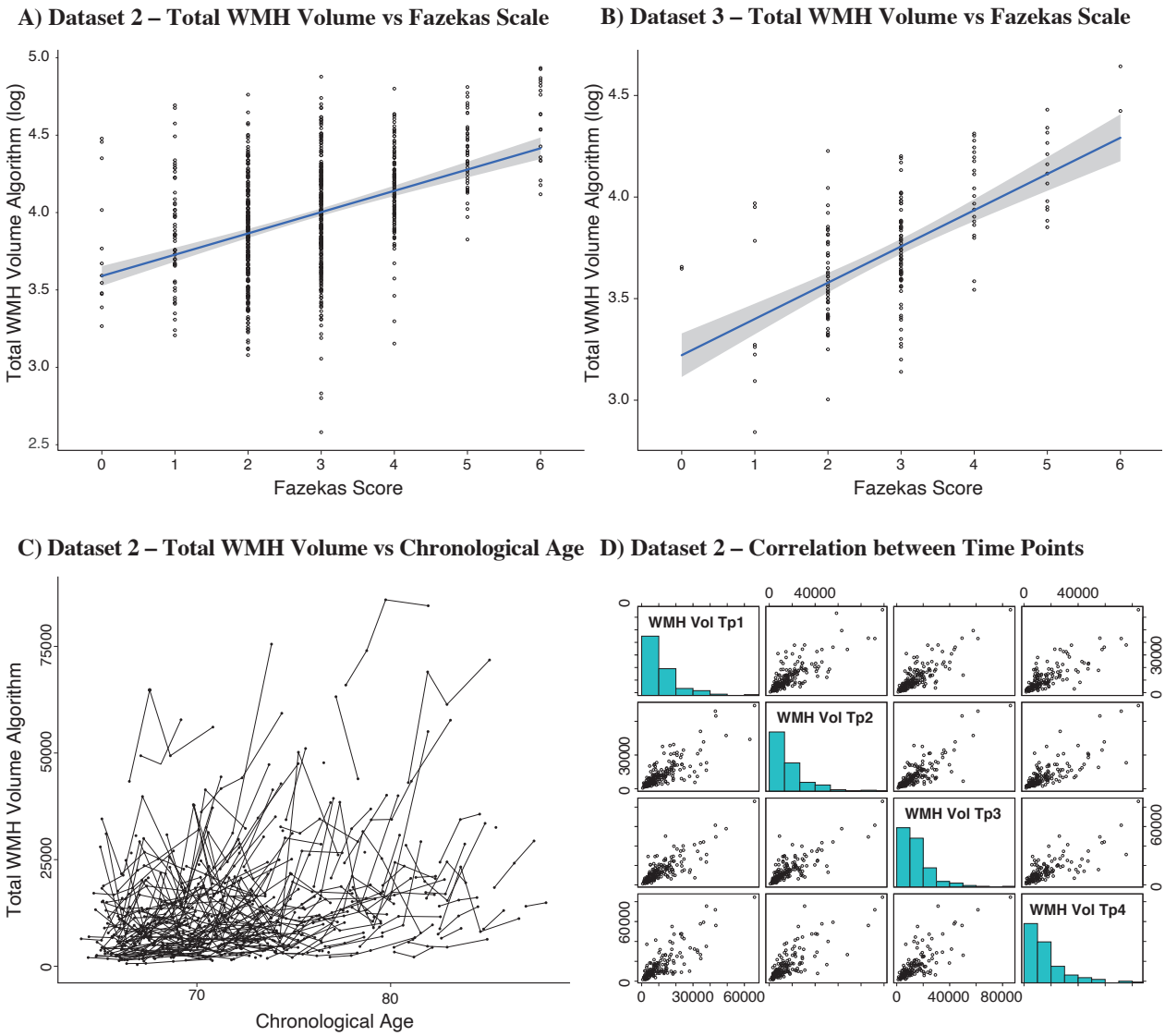
579 ***UBO validation***. *Scatter plots of the log-transformed total WMH volume distribution according to*
580 *Fazekas scale with dataset 2 and 3 (A and B), spaghetti plot of total WMH volume and chronological*
581 *age of dataset 2 (C), and the Pearson's product-moment correlation between the time points with*
582 *dataset 2 (D).*



583
584
585
586
587
588
589
590
591
592

593 **Figure 3**

594 ***BIANCA validation.*** *Scatter plot of the log-transformed total WMH volume distribution according to*
595 *Fazekas scale with dataset 2 and 3 (A and B), spaghetti plot of total WMH volume and chronological*
596 *age of dataset 2 (C), and the Pearson's product-moment correlation between the time points with*
597 *dataset 2 (D).*



598
599
600
601
602
603
604
605
606
607

Table 7

Summary of the results of the validation of the algorithms: Associations between total WMH volume, periventricular WMH volume, deep WMH volume and the ***Fazekas scores*** *(totalWMH – Fazekas, PVWMH – Fazekas, DWMH – Fazekas) and also for* ***chronological age*** *(totalWMH – Age) for each algorithm and input modality, and correlations* ***between the time points*** *(Tp1–Tp2, Tp2–Tp3, Tp3–Tp5. The highest $\hat{\beta}$ and correlation (r) values are marked in bold, the lowest $\hat{\beta}$ and correlations (r) values are characterized by gray shadowed cells.*

	FreeSurfer T1w ***N* = 800**	**UBO Detector 3D FLAIR + T1w** ***N* = 166**	**UBO Detector 2D FLAIR + T1w** ***N* = 756**	**BIANCA 3D FLAIR + T1w** ***N* = 166**	**BIANCA 2D FLAIR + T1w** ***N* = 762**
totalWMH – Fazekas[a]	$\hat{\beta}$ = 0.760***	$\hat{\beta}$ = 0.306***	**$\hat{\beta}$ = 0.800*** **	$\hat{\beta}$ = 0.639***	$\hat{\beta}$ = 0.419***
CI	0.715 – 0.805	0.162 – 0.450	0.757 – 0.843	0.515 – 0.716	0.356 – 0.483
Cohen's *d*[b]	1.805	1.629	2.228	1.038	0.514
PVWMH – Fazekas[a]	†	$\hat{\beta}$ = 0.706***	**$\hat{\beta}$ = 0.713*** **	$\hat{\beta}$ = 0.590***	$\hat{\beta}$ = 0.344***
CI	†	0.598 – 0.815	0.663 – 0.763	0.464 – 0.717	0.279 – 0.410
Cohen's *d*[b]	†	1.417	1.459	0.906	0.393
DWMH – Fazekas[a]	†	$\hat{\beta}$ = 0.567***	**$\hat{\beta}$ = 0.643*** **	$\hat{\beta}$ = 0.426***	$\hat{\beta}$ = 0.364***
CI	†	0.441 – 0.693	0.589 – 0.698	0.285 – 0.569	0.299 – 0.429
Cohen's *d*[b]	†	0.840	1.099	0.522	0.423
totalWMH – Age[a]	**$\hat{\beta}$ = 0.454*** **	$\hat{\beta}$ = 0.306***	$\hat{\beta}$ = 0.390***	$\hat{\beta}$ = 0.356***	$\hat{\beta}$ = 0.295***
CI	0.394 – 0.514	0.162 – 0.450	0.326 – 0.454	0.214 – 0.497	0.227 – 0.362
Cohen's *d*[b]	0.580	0.330	0.468	0.408	0.329
Tp1 – Tp2[c]	***r* = 0.997*** **	§	*r* = 0.990***	§	*r* = 0.802***
Tp2 – Tp3[c]	*r* = 0.994***	***r* = 0.998*** **	*r* = 0.992***	*r* = 0.716***	*r* = 0.807***
Tp3 – Tp5[c]	***r* = 0.995*** **	*r* = 0.970***	*r* = 0.983***	*r* = 0.917***	*r* = 0.768***

Notes: Confidence interval (CI), periventricular WMH (PVWMH), deep WMH (DWMH).

[a] $\hat{\beta}$ = fixed effect (slope). The WMH volumes are log-transformed and z-standardized, the Fazekas scores and the chronological age are z-standardized.

[b] Cohen's *d* = ≥ 0.2 – ≤ 0.5 = small = 0.5 – < 0.8 = moderate, 0.8 – ≥ 1.0 = large effect size (Cohen, 1988).

[c] Pearson's product-moment correlation coefficient = 0.1 – 0.3 = weak, 0.4 – 0.6 – 0.7 = moderate, 0.7 – 0.9 = strong, – 1 = perfect (Dancey & Reidy, 2017).

† Since FreeSurfer only reports total WMH volumes, no data is available for PVWMH and DWMH.

§ Due to insufficient longitudinal data, these correlations could not be calculated.

*** $p < 0.001$

3.3. Longitudinal post hoc analyses of the algorithms' output

Throughout the BIANCA outputs, we detected massive WMH volume fluctuations in the individual change trajectories over time that could not be seen in the cross-sectional data with the manual gold standard. An example of a segmentation of a BIANCA 2D FLAIR output over four time points of one subject is illustrated in **Supplement Figure 4** (total WMH volume of the different time points: Tp1 = 5.136 cm^3, Tp2 = 4.213 cm^3, Tp3 = 20.746 cm^3, Tp5 = 7.848 cm^3). In this example, Tp 3 shows a huge deviation. These WMH volume fluctuations between the time points are also reflected

visually in a suspicious zigzag pattern in **Figure 3,** panel C. Further, compared to the other algorithms, BIANCA showed lower association between the total WMH, PVWMH, and DWMH volumes and the Fazekas scores as well as lower correlations between the time points compared to FreeSurfer and UBO Detector (see **Table 7**). Since it is crucial to know whether these within-subject WMH volume fluctuations are only isolated cases or a more wide-spread problem, we evaluated the amount of suspicious fluctuations. For this purpose, we checked the entire WMH volume output files of BIANCA 3D FLAIR and examined randomly the 2D FLAIR WMH volume output files. We also reviewed the BIANCA segmentation masks of the suspicious values. It was found that all of these WMH segmentations masks contained erroneous voxels, particularly in the following areas: semi-oval centre, orbitofrontal cortex (orbital gyrus, gyrus rectus, above the putamen), and occipital lobe below the ventricles. Although recent studies indicate some WMH variability (Shi & Wardlaw, 2016), we assumed that these massive peaks are driven by erroneous segmentation of the algorithm. Since our knowledge of WMH volume changes (Shi & Wardlaw, 2016) is insufficient and has inconsistent findings (Ramirez, McNeely, Berezuk, Gao, & Black, 2016), our aim was to estimate plausible ranges, in which WMH volume increases and decreases can occur, based on the WMH volumes outputted by the different algorithms in our study.

In a first step, we determined how many comparisons between two measurement points were possible for each subject in the entire data set. We differentiated between 1-year intervals (for example: baseline – 1-year follow-up / 1-year follow-up – 2-year follow-up), 2-year-intervals, 3-year-intervals, and 4-year-intervals. For BIANCA 2D FLAIR, n = 209 subjects provide longitudinal imaging data (at least two time points). These 2D FLAIR images (N = 762 images) allowed 531 intervals between two measurement points (1-year intervals: n = 369, 2-year intervals: n = 145 etc.). Importantly, overlapping intervals were not included. If a subject had data for the baseline, 1-year and 2-year follow-up, only the comparisons «baseline vs. 1-year follow-up» and «1-year follow-up vs. 2-year follow-up» were considered. The comparison «baseline vs. 2-year follow-up» was only considered if a subject missed 1-year follow-up data.

Secondly, we calculated the percent WMH volume change from the previous to the subsequent measurement point for each subject, for each combination of algorithm and modality (e.g., BIANCA 2D FLAIR + T1w), and for the available time interval (1-year interval). Then, for each algorithm-by-modality combination and available time interval, the mean percent change was determined as well as the standard deviation (SD). These metrics were averaged across the five algorithm-by-modality combinations and used to calculate a general «range of tolerance», in which WMH volume increase and decrease for a given time interval were considered representing true change. The upper limit of the increases (average across algorithm-by-modality + 1SD) and the lower limit of the decreases (average across algorithm-by-modality - 1SD) were used to identify the suspicious cases.

Supplementary Table 9 lists the algorithm- and modality-specific WMH volume changes in percent, the mean WMH volume changes across all algorithms, and the associated «range of tolerance» for all time intervals. WMH volume increases and decreases outside of these «ranges of tolerance» were considered as suspicious. **Table 8** provides information on the number of suspicious changes per algorithm and modality in comparison to the number of possible comparisons between two measurement points. This table also contains information on the distribution of suspicious changes in dependence of WMH load (low, medium, high) – using the Fazekas scale – and informs about the percentage of subjects, in which suspicious WMH volume changes occurred. The categories were divided into the following categories: Fazekas score 0 – 2 = low WMH load, Fazekas score 3 and 4 = medium WMH load, and Fazekas score 5 and 6 = high WMH load. In each case, the previous measurement point, in the individual change trajectories over time, served as the basis for the classification.

Compared to the other algorithms, BIANCA (dataset 2 and 3) clearly shows the most «subjects with suspicious longitudinal data» (dataset 2: n = 109, 52.15%; 3D FLAIR: n = 7, 17.95%), and «suspicious intervals between two measurement points» (dataset 2: n = 161, 30.32%; dataset 3: n = 8, 16.67%), see **Table 8**. Furthermore, BIANCA (2D and 3D FLAIR) was the only algorithm that generated more «suspicious intervals between two measurement points» than «subjects with suspicious longitudinal data», indicating the zigzag pattern within the subject's trajectories.

Considering the WMH volume increases and decreases in all time intervals, BIANCA showed the highest percentages. In the 1-year intervals, for example, the 3D FLAIR image outputs showed a mean increase of 76.82% (SD = +174.65%; *n* = 9) and a mean decrease of -30.63% (SD = - 33.06%; *n* = 3). The 2D FLAIR image outputs of BIANCA revealed a mean increase of 64.80% (SD = +76.92%; *n* = 236) and a mean decrease of -26.61% (SD = -19.41%; *n* = 133), see **Supplementary Table 9** for more results. Regarding the WMH load, no systematic pattern could be detected within BIANCA. However, in BIANCA 2D FLAIR 55.90% of the «suspicious intervals between two measurement points» were images with a low WMH load, and 40.99% with a medium WMH load. With BIANCA 3D FLAIR, the suspicious estimates pertained to more images with a medium WMH load. However, there was much less data available (BIANCA: *n* = 8; UBO: *n* = 2; FreeSurfer: *n* = 0).

Table 8

Display per algorithm with the respective input modality (T1w, 2D FLAIR + T1w, 3D FLAIR + T1w) according the WMH load of the total number (N) of outputted images per algorithm, number (n) of subjects with longitudinal data (at least 2 time points), number and percentage (in brackets) of subjects with suspicious longitudinal data, number of intervals between two measurement points, and number and percentage (in brackets) of suspicious intervals between two measurement points.

Algorithm Modality WMH load	**N of outputted images**	**n of subjects with longitudinal data**	**n (and %) of subjects with suspicious longitudinal data**		**n of intervals[a] between two measurement points**	**n (and %) of suspicious intervals[a] between two measurement points**	
BIANCA 2D FLAIR	762	209	**109**	**(52.15%)**	**531**	**161**	**(30.32%)**
low						90	(55.90%)
medium						66	(40.99%)
high						5	(3.11%)
BIANCA 3D FLAIR	166	39	**7**	**(17.95%)**	**48**	**8**	**(16.67%)**
low						3	(37.50%)
medium						5	(62.50%)
high						0	(0%)
UBO 2D FLAIR*	757†	209	**7**	**(3.35%)**	523	**7**	**(1.34%)**
low						4	(57.14%)
medium						1	(14.29%)
high						2	(28.57%)
UBO 3D FLAIR	166	39	**2**	**(5.13%)**	48	**2**	**(4.17%)**
low						0	(0%)
medium						2	(100.00%)
high						0	(0%)
FreeSurfer T1w	800	213	**0**	**(0%)**	569	**0**	**(0%)**
low						0	(0%)
medium						0	(0%)
high						0	(0%)

Notes: WMH load is divided into low, medium, and high WMH load according to the Fazekas scale. Fazekas score 0 – 2 = low WMH load, Fazekas score 3 and 4 = medium WMH load, and Fazekas score 5 and 6 = high WMH load.
[a] Explanation «intervals between two measurement points»: If a subject had 3 time points (Tp1, Tp2 and Tp5) this would result in two existing intervals.
† The data point with the segmentation error (segmented eyeballs) is included (see 2.7.2 UBO Detector).

4. Discussion

In this study we compared and validated the performance of three freely available and automated algorithms for WMH extraction: FreeSurfer, UBO Detector, and BIANCA by applying a standardized protocol on a large longitudinal dataset of T1w, 3D FLAIR and 2D FLAIR images from cognitively healthy older adults with a relatively low WMH load. We discovered that all algorithms have certain strengths and limitations. FreeSurfer showed deficiencies particularly with respect to segmentation accuracy (i.e, DSC) and clearly underestimated the WMH volumes. We therefore argue that it cannot be considered as a valid substitution for manually segmented WMH. BIANCA and UBO Detector showed a higher segmentation accuracy compared to FreeSurfer. When using 3D FLAIR + T1w images as input, BIANCA performed significantly better than UBO Detector regarding the accuracy metrics DER and H95. However, we identified a significant amount of outlier WMH volumes estimations in the within-subject change trajectories of the BIANCA volume outputs. Exploratory analyses of these suspicious fluctuations indicated random false segmentations which contribute to the erroneous volume estimations. UBO Detector – as a fully automated algorithm – had the best cost-benefit ratio in our study. Although there is room for optimization regarding segmentation accuracy, it distinguished itself through its excellent volumetric agreement with the manually segmented WMH in both FLAIR modalities (as reflected by the ICCs) and its high associations with the Fazekas scores. In addition, it proved to be a robust estimator of WMH volumes over time.

4.1. Evaluation of the algorithms

FreeSurfer. The total WMH volumes from FreeSurfer showed a strong linear association with the Fazekas scores, and also a strong correlation between the time points. As expected, chronological age showed a lower association with the WMH volumes than with the Fazekas scores. It showed no

729 suspicious WMH volume increases and decreases between time points in the longitudinal data. These
730 results show that FreeSurfer outputs reliable data over time. However, we also observed that the
731 WMH volume validity of FreeSurfer's output is not given due to the fundamental underestimation of
732 WMH volume compared to the corresponding T1w gold standard. This can be attributed to the fact
733 that WMH often appear isointense in T1w sequences and are therefore not detected (Wardlaw et al.,
734 2013). Furthermore, the lower contrast of the DWMH compared to the PVWMH, which is due to the
735 lower water content in the DWMH as a result of the longer distance to the ventricles, might
736 contribute to the WMH volume underestimation. FreeSurfer often omitted DWMH, a finding also
737 reported by Olsson et al. (2013). In addition, our analyses showed that FreeSurfers' underestimation
738 of the WMH volume was even more pronounced in high WMH load images (see Bland-Altmann Plot
739 (Bland & Altman, 1986) in **Figure 4,** panel C). The same bias was shown by Olsson et al. (2013)
740 when comparing the volumes of the semi-manually segmented WMH (2D FLAIR) and the
741 FreeSurfer (T1w) output. One reason behind this WMH volume underestimation in subjects with
742 high WMH loads might be the fact that FreeSurfer segments them as grey matter (e. g. for the
743 bilateral caudate) (Dadar, Potvin, Camicioli, & Duchesne, 2020) which could also explain
744 FreeSurfer's low false positive ratio. The spatial overlap performance of FreeSurfer in our study is
745 comparable to the findings in the validation study reported by Samaille and colleagues (2012) with a
746 cohort of mild cognitive impairment and CADASIL patients. Smith and colleagues (2011) reported
747 an Intraclass Correlation Coefficient of 0.91 between Freesurfer's output and the WMH manually
748 segmented in ten images from one operator. Other accuracy metrics were not calculated. Ajilore et al.
749 (2014) reported a high correlation of $r = 0.91$ between the automated WMH volume of FreeSurfer
750 and the WMH volume of their manual segmentation in T1w and T2w images, including 20 subjects
751 with late-life major depression. However, they have not described the manual segmentation
752 procedure in detail. Nevertheless, this WMH volume underestimations between T1w and both
753 FLAIR modalities, is in line with the STandards for ReportIng Vascular changes on nEuroimaging
754 (STRIVE), stating that FLAIR images tend to be more sensitive to WMH and therefore are
755 considered more suitable for WMH detection than T1w images (Dadar et al., 2018; Wardlaw et al.,

2013). However, comparisons to previous studies are difficult because (a) for most studies it is not indicated whether they used a fully manually segmented gold standard or a gold standard generated by a semi-automated method, (b) sample sizes have been small, (c) there is very little information on sensitivity, H95, FPR, DER, OER and (d) FreeSurfer has not been applied to longitudinal data. Moreover, to our knowledge, previous studies have not compared FreeSurfer's WMH volumes with manual segmentations on T1w structural images or with visual rating scales such as the Fazekas scale. Although, in our study, FreeSurfer's WMH volumes are strongly associated with the Fazekas scores and showed reliable WMH volume increases and decreases over time. However, FreeSurfer cannot be considered as a valid substitute for manual WMH segmentation on this dataset due to the weak outcomes in the accuracy metrics (DSC, OER, ICC), and especially due to its massive WMH volume underestimation. Nevertheless, because of the valid and reliable outcomes with the Fazekas scale, FreeSurfer is suitable for use in clinical practice, even though its values should not be interpreted as absolute values.

UBO Detector. The total WMH volumes estimated by UBO Detector on the basis of 3D or 2D FLAIR input were associated with the Fazekas scores (strong effect) and with age (weak effect). Also, the PVWMH and DWMH volumes with both FLAIR (+ T1w) modalities as input showed a strong linear association with the Fazekas scores. This is consistent with the paper by the developers of the UBO Detector (Jiang et al., 2018) in which they reported significant associations between UBO Detector-derived PVWMH and DWMH volumes and Fazekas scale ratings. The results of their volumetric agreements – calculated with ICCs – were similar to our results, especially for dataset 2 (2D FLAIR + T1w). Further, the DSC and ICC values were similar to those of the very recent cross-sectional study of Vanderbecq et al. (2020). In our analysis, we found slightly higher correlations between the time points of our longitudinal datasets than Jiang et al. (2018), and a better false positive ratio. On the other hand, we were not able to replicate their high values based on 2D FLAIR images in sensitivity and the overlap measurements (DSC, DER, and OER). In their study based on their 2D FLAIR datasets – the H95 was not calculated, and no 3D FLAIR data was available for comparisons. The difference regarding the sensitivity and overlap measurements may be due to the

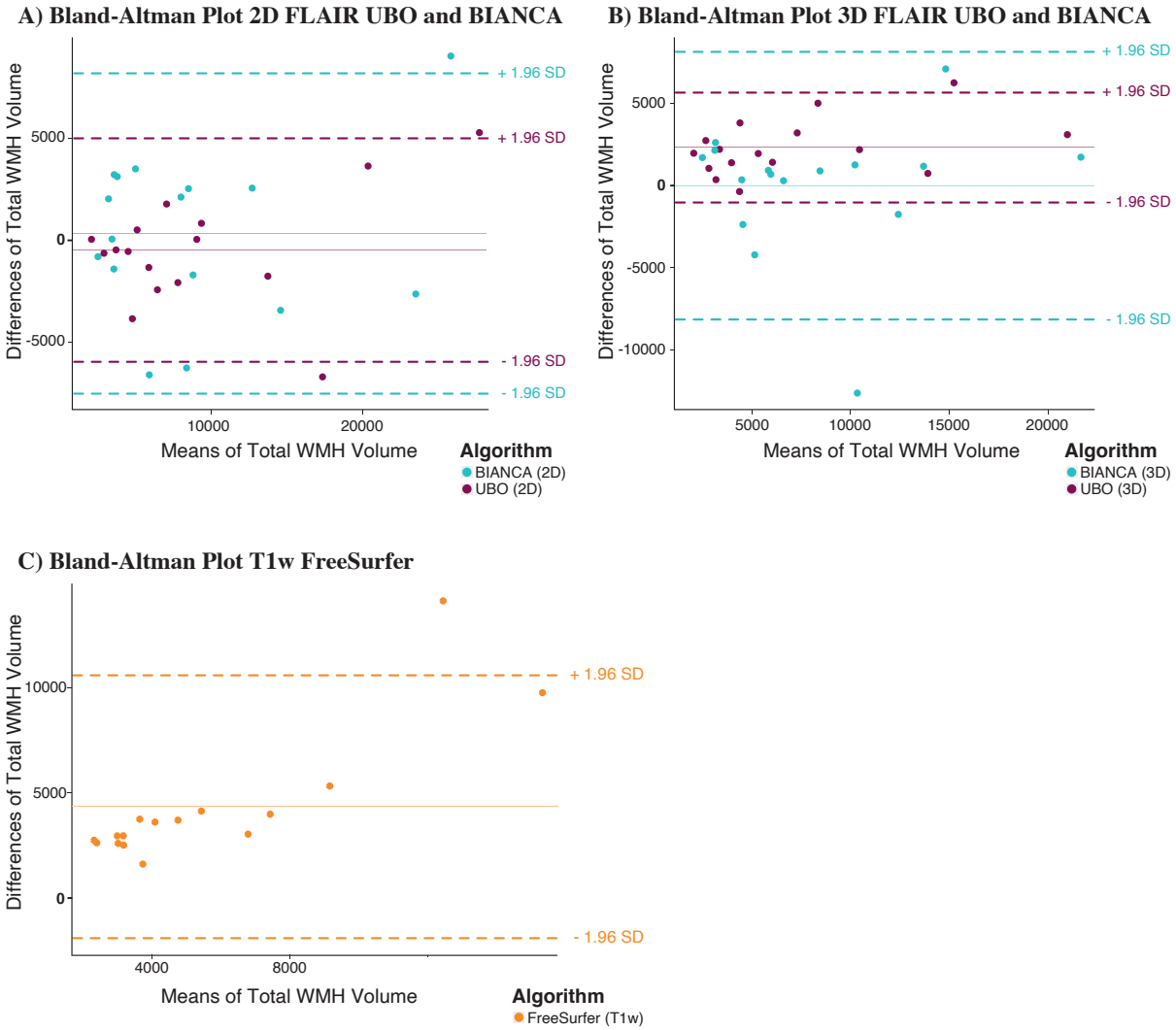
783 fact that, in contrast to Jiang et al. (2018), we did not use a customized training dataset but the 2D
784 FLAIR based built-in training dataset. Hence, whether a 2D or a 3D FLAIR image input is used for
785 UBO Detector may influence the estimated WMH volumes. Indeed, our analysis indicated that the
786 WMH volumes estimated by UBO Detector depends on the modality of the FLAIR input (2D vs. 3D
787 FLAIR). Volumes extracted from dataset 2 (2D FLAIR + T1w images) tended to be more similar to
788 the WMH volume of the gold standard with the same modality, while volumes extracted from dataset
789 3 (3D FLAIR + T1w mages) tended to underestimate the WMH volume of the respective gold
790 standard (see **Table 5,** and Bland-Altmann Plot in **Figure 4,** panel A and B). For several reasons
791 UBO Detector's longitudinal pipeline was not used in our study. First, UBO Detector requires an
792 equal number of sessions for all subjects, which would have resulted in a reduction of our sample
793 size. Secondly, it registers all sessions to the first time point, an approach which has been shown to
794 lead to biased registration (Reuter, Schmansky, Rosas, & Fischl, 2012). Lastly, comparing the two
795 pipelines, Jiang et al. (2018) did not find significant differences regarding the extracted WMH
796 volumes. Up to now, we have not found any other study that validated UBO Detector and/or
797 compared it with other WMH extraction methods on a longitudinal dataset with different MR
798 modalities.
799 **BIANCA.** In order to properly compare our output data from BIANCA with the results of the
800 original study from BIANCA we additionally run BIANCA's evaluation script to calculate the same
801 metrics (**Supplementary Table 7**). Our overall results for the different accuracy metrics
802 corresponded more to those of their vascular cohort than to those of their neurodegenerative cohort.
803 We received quite similar results for the 2D FLAIR modalities with respect to the correlations
804 between the BIANCA WMH volumes and our manually segmented WMH volumes. However, we
805 were not able to replicate the high ICCs and the high associations between the WMH volume and the
806 Fazekas scores they received in both cohorts, neither for dataset 2 nor for dataset 3 images as input
807 modalities. With our dataset 2 BIANCA showed only a moderate linear association with the Fazekas
808 scale and the total WMH volume, and a weak association for PVWMH and DWMH – although a
809 customized training dataset was available. We also obtained similarly small associations for WMH

810 volumes and age for both datasets as Griffanti et al. (2016) did for their neurodegenerative cohort.
811 We suspect that this might be due to the false segmentations in our BIANCA outputs. While the
812 effect of false WMH segmentations was not detected in the cross-sectional analyses, it was
813 uncovered because of massive WMH volume fluctuations in the within-subjects trajectories. The
814 developers of UBO Detector compared their algorithm also to BIANCA – based on a sample of 40
815 subjects, and noticed that BIANCA tended to overestimate the WMH in «milky» regions, whereas
816 the sensitivity for WMH detection was higher in BIANCA than in UBO Detector, which is in line
817 with our findings.
818 BIANCA features **LOCATE** as a method to determine spatially adaptive thresholds in different
819 regions in the lesion probability map. Sundaresan et al. (2018) showed that LOCATE is beneficial
820 when the BIANCA algorithm is trained with dataset-specific images or when the training dataset was
821 acquired with the same sequence and the same scanner. For the group of healthy controls, they
822 achieved similar visual outputs with LOCATE as compared to those with global thresholding.
823 However, since no manually segmented gold standard was available for the healthy controls in their
824 study, a quantitative comparison between manually segmented WMH and LOCATE's WMH was not
825 possible. In our analysis, LOCATE, as compared to BIANCA's global thresholding did perform
826 significantly worse at processing images with a low WMH load (see **Supplementary Analysis**).
827 LOCATE arguably had more true positives, resulting in significantly higher sensitivity, but this came
828 at the cost of a 3-fold higher false positive rate (see **Supplementary Table 10**). Hence, all other
829 metrics (DSC, OER, DER, H95 and FPR) showed worse outcomes for LOCATE compared to
830 BIANCA's global thresholding in our dataset. In addition, the WMH volumes received from
831 LOCATE deviated significantly from the WMH volumes of the gold standards which is due to the
832 massive number of false positives in LOCATE (see **Supplementary Figure 5**). Ling and colleagues
833 (2018) validated BIANCA with different input modalities (single FLAIR or FLAIR + T1w), with a
834 cohort of patients with CADASIL using a semi-manually generated gold standard of ten subjects per
835 modality. In their dataset, which contained an extremely high WMH load, they received a median
836 DSC for the 2D FLAIR + T1w images of 0.79 (our median 2D FLAIR = 0.560) and a median DSC

for the 3D FLAIR + T1w images of 0.76 (our median DSC 3D FLAIR = 0.615). The volumetric agreement, measured by the ICC, for their 3D FLAIR + T1w, and 2D FLAIR + T1w with a mixed WMH load for the training data set, were very similar to ours with the global threshold method. In the study by Vanderbecq et al. (2020), the ICC determined with their «clinical routine data set» (patients who were referred for assessment of cognitive impairment) is comparable to ours with the 2D FLAIR input, whereas the ICC determined with their «research data set» (ADNI dataset; comprising mainly Alzheimer's and amnestic mild cognitive impairment patients) was lower compared to ours. Ling et al. (2018) found that BIANCA tended to overestimate the WMH volumes in subjects with a low WMH load and underestimate it in subjects with a high WMH load. According to them, in a group of healthy elderly people with a low WMH exposure, such a bias would be unlikely to be identified. In both BIANCA outputs (3D and 2D FLAIR) we did not detect systematic biases, but revealed one clear underestimation in the subject with the highest WMH load in the 2D FLAIR images, and one pronounced overestimation in a subject with medium WMH load in the 3D FLAIR images (see Bland-Altman Plot **Figure 4,** panel A and B). With a similar approach, using the absolute mean WMH volume differences to the gold standards, we were able to show that the mean WMH volume differences of the WMH volumes of BIANCA are the results of random averaging over inaccurately estimated WMH volumes (see **Supplementary Table 8**). Given, that the focus of our study was to compare different algorithms in terms of costs and benefits, we did not test other settings for BIANCA but adhered to the default settings suggested in the original BIANCA validation. To our knowledge, BIANCA and LOCATE have not been validated with a longitudinal dataset so far.

Figure 4

Bland-Altman (Bland & Altman, 1986) *plots for WMH volume for the different algorithms (total WMH volume gold standard minus total WMH volume algorithm in mm³). The x-axes contain mean WMH volumes, the y-axes contain absolute differences of WMH volumes: S (x, y) = (gold standard WMH volumes (S1) + algorithm WMH volumes (S2)/2, S1 - S2).*



4.2. Comparison of the algorithms with each other

The quality assessment of the algorithms is critically based on manually segmented WMH, here referred as gold standards. In order to prove construct validity, 16 gold standards per modality (T1w, 3D FLAIR, 2D FLAIR) were correlated amongst each other and with the respective WMH volume outcomes of the algorithms (**Supplementary Figure 2**). The WMH volumes of the manually segmented WMH correlated very strongly (all combinations: $r = 0.97$, $p < 0.05$), indicating a very high validity for our gold standard. UBO Detector applied to dataset 3 showed the highest

correlations with the gold standard WMH volumes, followed by UBO Detector applied to dataset 2, FreeSurfer applied to dataset 1, BIANCA applied to dataset 2, and BIANCA applied to dataset 3. This correlation pattern is interesting and partly unexpected, especially when considering that BIANCA was the only algorithm that was fed with a customized training dataset for every modality (3D FLAIR and 2D FLAIR images trainings dataset). UBO Detector, on the other hand, has a built-in training dataset based on 2D FLAIR images which was used for the analyses of both FLAIR input modalities – 3D and 2D FLAIR images. Remarkably, the 3D and 2D FLAIR images based WMH volumes of UBO Detector correlated strongly with the respective gold standards. Across modalities, the correlation of the algorithm's WMH volume outputs was very high between UBO 3D and 2D FLAIR images ($r = 0.99$), while the correlation between BIANCA 3D and 2D FLAIR images was clearly smaller ($r = 0.68$). In the study of Vanderbecq et al. (2020), UBO Detector showed better results than BIANCA in the WMH volume evaluations (ICC and volume similarity), and for BIANCA the DSC was significantly smaller in images with artifacts compared to images without artifacts, what might be reflected by these worse WMH volume evaluations – this was not the case for UBO Detector. Interestingly, the WMH volumes of FreeSurfer correlated also very high with the two UBO Detector modality outputs (2D FLAIR + T1w: $r = 0.92$, 3D: $r = 0.949$), but less strong with the two BIANCA modality outputs (2D FLAIR + T1w: $r = 0.75$, 3D: $r = 0.86$). In summary, these correlations show that the WMH volume segmentations of UBO Detector and FreeSurfer are better aligned than those of BIANCA, which could indicate the WMH segmentation errors we had seen in BIANCA. On the other hand, the WMH volumes extracted by FreeSurfer were generally smaller than the outputs of UBO Detector and BIANCA. We would like to emphasize that FreeSurfer is the only algorithm that has even underestimated its «own» gold standard. This WMH volume underestimation also affected the accuracy metrics (DSC, Sensitivity, OER and ICC), which were significantly worse for FreeSurfer compared to the other algorithms. Regarding the H95 and the DER, BIANCA 3D FLAIR performed best. On the flip side BIANCA showed the weakest association of WMH volume and Fazekas scores, and the smallest correlation of WMH volumes between all time points. In line with the latter, BIANCA with the 3D FLAIR + T1w and 2D FLAIR + T1w images as input showed

906 the most «suspicious intervals between two measurement points» and «subjects with suspicious
907 longitudinal data» compared to the other algorithms, and also as the only algorithm more «suspicious
908 intervals between two measurement points» than «subjects with suspicious longitudinal data». These
909 differences between intervals and subjects reflect the previously discovered visual zigzag pattern in
910 trajectories within subjects. The strongest WMH increases over time in literature were observed in
911 subjects with a high WMH at baseline (Duering et al., 2013; Gouw et al. 2008; R. Schmidt et al.,
912 2003), and little progression in punctate WMH but rapid progression in confluent WMH are reported
913 (Schmidt, Seiler, & Loitfelder, 2016). However WMH can also cavitate to take on the appearance of
914 lacunes, and so they can also disappear (Shi & Wardlaw, 2016) for example after therapeutic
915 intervention (Ramirez et al., 2016). Some studies report annual percentage increases in WMH
916 volumes in the range between 12.5% and 14.4% in subjects with early confluent lesions, and 17.3%
917 and 25.0% in subjects with confluent abnormalities (Duering et al., 2013; Sachdev, Wen, Chen, &
918 Brodaty, 2007; Schmidt et al., 2003; van Dijk et al., 2008). Ramirez et al. (2016) summarized the
919 progression rates of WMH volumes in serial MRI studies in their Table 2, showing a wide variability
920 of ranges. BIANCA applied to dataset 3 showed an annual WMH volume increase range of 76.8%
921 (mean) to 251.5% (mean + 1SD), and a range of 64.8% to 141.7% when applied to dataset 2.
922 Comparing this progression of WMH volume growth with those described in the literature, and also
923 considering that the WMH in healthy older subjects should increase only slightly, then these results
924 again may point to segmentation errors, that influenced the segmentation's reliability. Having a
925 closer look on the segmentation variability of the algorithms by means of the Bland-Altman plots
926 **(see Figure 4)**, we observed that the limits of agreement are wider in BIANCA than in UBO
927 Detector. Moreover, the 2D and 3D FLAIR image plots show strong outliers (under- and
928 overestimations). Interestingly, in the same 16 subjects used for the gold standard, the single
929 deviations in BIANCA's output seem to cancel each other out and result in a mean WMH volume
930 that is very similar to the gold standard's WMH volume (see **Supplementary Table 8**). Regardless
931 of the WMH load, UBO Detector systematically underestimated the WMH volumes in the 3D FLAIR
932 images compared to the gold standard, and BIANCA also showed a slight tendency to do so (see

933 **Figure 4,** panel B). This effect, however, was not significant (**see Table 6**). From analyzing the
934 validation of the algorithms, we can conclude that with our datasets UBO Detector and FreeSurfer, as
935 compared to BIANCA, performed more robust and also more consistent across time. Future research
936 needs to evaluate if the segmentation errors, BIANCA produced with our longitudinal dataset, occur
937 also in the context of other datasets.
938 One general problem in the context of automated WMH lesion segmentation using FLAIR images is
939 the incorrect inclusion of the septum pellucidum, the area separating the two lateral ventricles, in the
940 output masks. This area appears hyperintense on FLAIR sequences, and therefore, looks very similar
941 to WMH. When erroneously detected as WMH, the septum pellucidum enters the output volumes as
942 false positive region, which leads to an overestimation of the WMH volume. The UBO Detector
943 developers also segmented the septum pellucidum in their gold standard (see their Supplementary
944 Figure 1b). Since they already fed their algorithm with this false positive information, it was to be
945 expected that UBO Detector would also segment the septum pellucidum in our data, which may have
946 caused the worse DER compared to FreeSurfer and BIANCA. Interestingly, however, also the
947 BIANCA and LOCATE outputs included some false positives in the septum pellucidum although the
948 algorithm was trained with a customized training dataset in which this area was not segmented.

949 4.3. Strength and limitations

950 The main strengths of this study are the validation and comparison of three freely available
951 algorithms using a large longitudinal dataset of cognitively healthy subjects with a low WMH load.
952 Furthermore, we used fully manually gold standards, segmented in all three planes (sagittal, coronal,
953 axial), and for three different MR modalities (T1w, 3D FLAIR, 2D FLAIR). For the 3D and 2D
954 FLAIR modalities, three operators segmented the same images, reaching excellent inter-operator
955 agreements (Caligiuri et al., 2015; Dadar et al., 2017). Supporting the high quality of our manual
956 segmentations, we found an excellent (Cicchetti, 1994) reliability for the total volumetric agreement
957 between the measurements of the 3D and 2D FLAIR images (ICC: 3D FLAIR mean = 0.96; 2D
958 FLAIR = 0.82). There were also no significant mean WMH volume differences and high correlations

959 within the gold standards. Besides the manual segmentations, the whole dataset (N = 800) was rated
960 by the three operators using the Fazekas scale, so that an inter-operator reliability overall and
961 between the three operators could be calculated. The rating comparisons of all three operators
962 resulted in substantial to almost perfect agreement (Landis & Koch, 1977). Thus, the Fazekas scale
963 could be considered a reliable gold standard for cross-validation of WMH volumes extracted with the
964 automatic algorithms.
965 The limitation of this study is that it is only one sample with homogeneous participants, which on the
966 other hand makes the data comparable. Future studies will determine how well these results
967 generalize to other studies, scanners, sequences, and heterogeneous datasets with clinical participants.

968 **4.4. Usability of the algorithms**

969 Given that the algorithms for WMH extraction are usually not implemented by trained programmers,
970 usability is an important issue to mention here.
971 FreeSurfer has not been specifically programmed for WMH detection, but is a tool for extensive
972 analysis of brain imaging data. Because of all the other parameters FreeSurfer outputs besides WMH
973 volume, the processing time is very long (many hours per session). The FreeSurfer output comprises
974 the total WMH volume and the total non-WMH volumes (grey matter).
975 UBO Detector has been specifically programmed for WMH detection, and has been trained with a
976 «built-in» training dataset. Theoretically, it is possible to train the algorithm using a previously
977 manually segmented gold standard. However, this procedure does only work within the Graphic User
978 Interface (GUI) in DARTEL space, and is very time-consuming. The output from UBO Detector is
979 well structured and contains among others the WMH volume and the number of clusters for total
980 WMH, PVWMH, DWMH as well as WMH volumes per cerebral lobe. For subset 2 the WMH
981 extraction for the whole process – incl. pre- and postprocessing – took approx. 14 min per brain with
982 the computing environment specified in the methods section. For subset 3 the WMH extraction took
983 approx. 32 min per brain.

BIANCA is a tool integrated in FSL (FMRIB's Software Library) with no need of any other program. It is very flexible in terms of MRI input modalities that can be used and offers many different options for optimization. The output of BIANCA comprises the total WMH. If required, a distance from the ventricles can be selected to PVWMH and DWMH. In a longitudinally study with many subjects and time points, or also in a study with a big sample size, the aggregation of the algorithm output files seemed to be very time-consuming because of the many output files. Our preprocessing steps for BIANCA took about 2:40 h per subject for the preparation of the templates, about 1:10 h per session for the preparation of the T1w, 2D and 3D FLAIR images. BIANCA required about 1:20 h per session for setting the threshold for both FLAIR images, and the WMH segmentation took about 4 min and 8 min per session for the 2D FLAIR and 3D FLAIR images, respectively.

5. Conclusions

The main aim of the current study has been the comparison and validation of three freely available algorithms for automated WMH segmentation using a large longitudinal dataset of cognitively healthy subjects with a relatively low WMH load. Our results indicate that FreeSurfer underestimates the total WMH volumes significantly and misses some DWMH completely. Therefore, this algorithm seems not suitable for research specifically focusing on WMH and its associated pathologies. However, since the high associations with the Fazekas scores, and the longitudinal robustness, could also indicate more importance in clinical practice than a high DSC with manual segmentations. BIANCA received in general good results in cross-sectional accuracy metrics but the longitudinal data suggest that it generates false segmented WMH volumes – frequently in frontal brain areas. UBO Detector, as a completely automated algorithm, has scored best regarding the costs and benefits due to its fully generalizable performance. Although UBO Detector performed best in this study in summary, improvements in accuracy metrics, such as DER and H95, would be desirable to be considered as a true replacement for manual segmentation of WMH. In general, this study shows how important the longitudinal validation of algorithms for the extraction of WMH can be – especially in

studies with many subjects, where not every image can be controlled by the human eye – in order to be able to detect suspicious variances at all.

6. CRediT author statement

Isabel Hotz: Conceptualization, Software, Methodology, Validation, Formal analysis, Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Pascal F. Deschwanden:** Methodology, Validation, Formal analysis, Data curation, Writing - Review & Editing. **Franz Liem:** Software, Formal analysis, Investigation, Data curation, Writing - Review & Editing. **Susan Mérillat:** Conceptualization, Investigation, Resources, Data curation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Spyros Kollias:** Validation, Supervision. Lutz Jäncke: Conceptualization, Investigation, Resources, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

7. Acknowledgement

We acknowledge all participants. We are thankful to Susanne Wehrli for her help on segmenting WMH masks used in our study, and Christoph Eberle for his valuable help with the graphics.
The current analysis incorporates data from the Longitudinal Healthy Aging Brain (LHAB) database project carried out at the University of Zurich (UZH). The following researchers at the UZH were involved in the design, set-up, maintenance and support of the LHAB database: Anne Eschen, Lutz Jäncke, Franz Liem, Mike Martin, Susan Mérillat, Christina Röcke, and Jacqueline Zöllig.

References

Ajilore, O., Lamar, M., Leow, A., Zhang, A., Yang, S., & Kumar, A. (2014). Graph Theory Analysis of Cortical-Subcortical Networks in Late-Life Depression. *The American Journal of Geriatric Psychiatry*, *22*(2), 195–206. doi:10.1016/j.jagp.2013.03.005

Anbeek, P., Vincken, K. L., van Osch, M. J. P., Bisschops, R. H. C., & van der Grond, J. (2004). Probabilistic segmentation of white matter lesions in MR imaging. *Neuroimage*, *21*(3), 1037–1044. doi:10.1016/j.neuroimage.2003.10.012

Baker, J. G., Williams, A. J., Ionita, C. C., Lee-Kwen, P., Ching, M., & Miletich, R. S. (2012). Cerebral small vessel disease: cognition, mood, daily functioning, and imaging findings from a small pilot sample. *Dementia and Geriatric Cognitive Disorders Extra*, *2*, 169–179. doi:10.1159/000333482

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Beauchemin, M., Thomson, K. P. B., & Edwards, G. (1998). On the hausdorff distance used for the evaluation of segmentation results. *Canadian Journal of Remote Sensing*, *24*(1), 3–8. doi:10.1080/07038992.1998.10874685

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *1*(8476), 307–310. doi:10.1016/S0140-6736(86)90837-8

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. doi:10.5334/joc.10

Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., & Cherubini, A. (2015). Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review. *Neuroinformatics*, *13*(3), 261–276. doi:10.1007/s12021-015-9260-y

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–

290. doi:10.1037/1040-3590.6.4.284

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. doi:10.1037/h0026256

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey, NJ: Lawrence Erlbaum Associates. doi:10.4324/9780203771587

Dadar, M., Maranzano, J., Ducharme, S., Carmichael, O. T., Decarli, C., Collins, D. L., & Alzheimer's Disease Neuroimaging Initiative. (2018). Validation of T1w-based segmentations of white matter hyperintensity volumes in large-scale datasets of aging. *Human Brain Mapping*, *39*(3), 1093–1107. doi:10.1002/hbm.23894

Dadar, M., Maranzano, J., Misquitta, K., Anor, C. J., Fonov, V. S., Tartaglia, M. C., … Alzheimer's Disease Neuroimaging Initiative. (2017). Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. *Neuroimage*, *157*, 233–249. doi:10.1016/j.neuroimage.2017.06.009

Dadar, M., Potvin, O., Camicioli, R., & Duchesne, S. (2020). Beware of white matter hyperintensities causing systematic errors in grey matter segmentations! *BioRxiv*. doi:10.1101/2020.07.07.191809

Dancey, C. P., & Reidy, J. (2017). Statistics Without Maths for Psychology. Retrieved July 2, 2020, from https://www.pearson.com/uk/educators/higher-education-educators/program/Dancey-Statistics-Without-Maths-for-Psychology-7th-Edition/PGM1768952.html

de Sitter, A., Steenwijk, M. D., Ruet, A., Versteeg, A., Liu, Y., van Schijndel, R. A., … MAGNIMS study group and for neuGRID. (2017). Performance of five research-domain automated WM lesion segmentation methods in a multi-center MS study. *Neuroimage*, *163*, 106–114. doi:10.1016/j.neuroimage.2017.09.011

DeCarli, C., Massaro, J., Harvey, D., Hald, J., Tullberg, M., Au, R., … Wolf, P. A. (2005). Measures of brain morphology and infarction in the framingham heart study: establishing what is normal. *Neurobiology of Aging*, *26*(4), 491–510. doi:10.1016/j.neurobiolaging.2004.05.004

Dubuisson, M. P., & Jain, A. K. (1994). A modified Hausdorff distance for object matching. In

Proceedings of 12th International Conference on Pattern Recognition (pp. 566–568). IEEE Comput. Soc. Press. doi:10.1109/ICPR.1994.576361

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., … Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. doi:10.1038/s41592-018-0235-4

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., & Zimmerman, R. A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Roentgenology*, *149*(2), 351–356. doi:10.2214/ajr.149.2.351

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*(2), 774–781. doi:10.1016/j.neuroimage.2012.01.021

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. doi:10.1016/s0896-6273(02)00569-x

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. doi:10.1016/0022-3956(75)90026-6

Frey, B. M., Petersen, M., Mayer, C., Schulz, M., Cheng, B., & Thomalla, G. (2019). Characterization of White Matter Hyperintensities in Large-Scale MRI-Studies. *Frontiers in Neurology*, *10*, 238. doi:10.3389/fneur.2019.00238

Gorgolewski, K. J., Alfaro-Almagro, F., Auer, T., Bellec, P., Capotă, M., Chakravarty, M. M., … Poldrack, R. A. (2017). BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods. *PLoS Computational Biology*, *13*(3), e1005209. doi:10.1371/journal.pcbi.1005209

Gorgolewski, K. J., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. doi:10.3389/fninf.2011.00013

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., … Jenkinson, M. (2016).

BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *Neuroimage*, *141*, 191–205. doi:10.1016/j.neuroimage.2016.07.018

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850–863. doi:10.1109/34.232073

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156. doi:10.1016/s1361-8415(01)00036-6

Jiang, J., Liu, T., Zhu, W., Koncz, R., Liu, H., Lee, T., … Wen, W. (2018). UBO Detector - A cluster-based, fully automated pipeline for extracting white matter hyperintensities. *Neuroimage*, *174*, 539–549. doi:10.1016/j.neuroimage.2018.03.050

Klöppel, S., Abdulkadir, A., Hadjidemetriou, S., Issleib, S., Frings, L., Thanh, T. N., … Ronneberger, O. (2011). A comparison of different automated methods for the detection of white matter lesions in MRI data. *Neuroimage*, *57*(2), 416–422. doi:10.1016/j.neuroimage.2011.04.053

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. doi:10.1016/j.jcm.2016.02.012

Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., … Biessels, G. J. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, *38*(11), 2556–2568. doi:10.1109/TMI.2019.2905770

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi:10.2307/2529310

Ling, Y., Jouvent, E., Cousyn, L., Chabriat, H., & De Guio, F. (2018). Validation and optimization of BIANCA for the segmentation of extensive white matter hyperintensities. *Neuroinformatics*, *16*(2), 269–281. doi:10.1007/s12021-018-9372-2

Mäntylä, R., Erkinjuntti, T., Salonen, O., Aronen, H. J., Peltonen, T., Pohjasvaara, T., & Standertskjöld-Nordenstam, C. G. (1997). Variable agreement between visual rating scales for white matter hyperintensities on MRI. Comparison of 13 rating scales in a poststroke cohort. *Stroke*, *28*(8), 1614–1623.

McCarthy, P. (2018). FSLeyes. *Zenodo*. doi:10.5281/zenodo.1887737

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. doi:10.1037/1082-989X.1.1.30

Moslem, S., Ghorbanzadeh, O., Blaschke, T., & Duleba, S. (2019). Analysing Stakeholder Consensus for a Sustainable Transport Development Decision by the Fuzzy AHP and Interval AHP. *Sustainability*.

Olsson, E., Klasson, N., Berge, J., Eckerström, C., Edman, A., Malmgren, H., & Wallin, A. (2013). White Matter Lesion Assessment in Patients with Cognitive Impairment and Healthy Controls: Reliability Comparisons between Visual Rating, a Manual, and an Automatic Volumetrical MRI Method-The Gothenburg MCI Study. *Journal of Aging Research*, *2013*, 198471. doi:10.1155/2013/198471

R Core Team, R. F. for S. C. (2020). R: A language and environment for statistical computing (4.0.4). Computer software, Vienna: Austria.

Ramirez, J., McNeely, A. A., Berezuk, C., Gao, F., & Black, S. E. (2016). Dynamic Progression of White Matter Hyperintensities in Alzheimer's Disease and Normal Aging: Results from the Sunnybrook Dementia Study. *Frontiers in Aging Neuroscience*, *8*, 62. doi:10.3389/fnagi.2016.00062

Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, *61*(4), 1402–1418. doi:10.1016/j.neuroimage.2012.02.084

Samaille, T., Fillon, L., Cuingnet, R., Jouvent, E., Chabriat, H., Dormont, D., … Chupin, M. (2012). Contrast-based fully automatic segmentation of white matter hyperintensities: method and validation. *Plos One*, *7*(11), e48953. doi:10.1371/journal.pone.0048953

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. doi:10.22237/jmasm/1257035100

Scheltens, P., Barkhof, F., Leys, D., Pruvo, J. P., Nauta, J. J., Vermersch, P., … Valk, J. (1993). A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. *Journal of the Neurological Sciences*, *114*(1), 7–12. doi:10.1016/0022-510x(93)90041-v

Schmidt, R., Seiler, S., & Loitfelder, M. (2016). Longitudinal change of small-vessel disease-related brain abnormalities. *Journal of Cerebral Blood Flow and Metabolism*, *36*(1), 26–39. doi:10.1038/jcbfm.2015.72

Shi, Y., & Wardlaw, J. M. (2016). Update on cerebral small vessel disease: a dynamic whole-brain disease. *Stroke and Vascular Neurology*, *1*(3), 83–92. doi:10.1136/svn-2016-000035

Shonkwiler, R. (1991). Computing the Hausdorff set distance in linear time for any Lp point distance. *Information Processing Letters*, *38*(4), 201–207. doi:10.1016/0020-0190(91)90101-M

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi:10.1037//0033-2909.86.2.420

Smith, E., Salat, D. H., Jeng, J., McCreary, C. R., Fischl, B., Schmahmann, J. D., … Greenberg, S. M. (2011). Correlations between MRI white matter lesion location and executive function and episodic memory. *Neurology*, *76*(17), 1492–1499. doi:10.1212/WNL.0b013e318217e7c8

Steenwijk, M. D., Pouwels, P. J. W., Daams, M., van Dalen, J. W., Caan, M. W. A., Richard, E., … Vrenken, H. (2013). Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage. Clinical*, *3*, 462–469. doi:10.1016/j.nicl.2013.10.003

Sundaresan, V., Zamboni, G., Le Heron, C., M. Rothwell, P., Husain, M., Battaglini, M., … Griffanti, L. (2018). Automated lesion segmentation with BIANCA: impact of population-level features, classification algorithm and locally adaptive thresholding. *BioRxiv*. doi:10.1101/437608

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C.

(2010). N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320. doi:10.1109/TMI.2010.2046908

van den Heuvel, D. M. J., ten Dam, V. H., de Craen, A. J. M., Admiraal-Behloul, F., van Es, A. C. G. M., Palm, W. M., … PROSPER Study Group. (2006). Measuring longitudinal white matter changes: comparison of a visual rating scale with a volumetric measurement. *American Journal of Neuroradiology*, *27*(4), 875–878.

van den Heuvel, T. L. A., van der Eerden, A. W., Manniesing, R., Ghafoorian, M., Tan, T., Andriessen, T. M. J. C., … Platel, B. (2016). Automated detection of cerebral microbleeds in patients with Traumatic Brain Injury. *NeuroImage. Clinical*, *12*, 241–251. doi:10.1016/j.nicl.2016.07.002

Vanderbecq, Q., Xu, E., Ströer, S., Couvy-Duchesne, B., Diaz Melo, M., Dormont, D., … Alzheimer's Disease Neuroimaging Initiative. (2020). Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. *NeuroImage. Clinical*, *27*, 102357. doi:10.1016/j.nicl.2020.102357

Wack, D. S., Dwyer, M. G., Bergsland, N., Di Perri, C., Ranza, L., Hussein, S., … Zivadinov, R. (2012). Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates. *BMC Medical Imaging*, *12*, 17. doi:10.1186/1471-2342-12-17

Wahlund, L. O., Barkhof, F., Fazekas, F., Bronge, L., Augustin, M., Sjögren, M., … European Task Force on Age-Related White Matter Changes. (2001). A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke*, *32*(6), 1318–1322. doi:10.1161/01.STR.32.6.1318

Wardlaw, J. M., Smith, E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., … STandards for ReportIng Vascular changes on nEuroimaging (STRIVE v1). (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurology*, *12*(8), 822–838. doi:10.1016/S1474-4422(13)70124-8

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments

in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. doi:10.1037/xge0000014