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

Corresponding Authors

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

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

Abstract

White matter hyperintensities of presumed vascular origin (WMH) are frequently found in MRIs of patients with various neurological and vascular disorders, but also in healthy elderly subjects. Although automated methods have been developed to replace the challenging task of manually segmenting the WMH, there is still no consensus on which validated algorithm(s) should be used. In this study, we validated and compared three freely available methods for WMH extraction: FreeSurfer, UBO Detector, and the Brain Intensity AbNormality Classification Algorithm, BIANCA (with the two thresholding options: global thresholding vs. LOCally Adaptive Threshold Estimation (LOCATE)) using a standardized protocol.

We applied the algorithms to longitudinal MRI data (2D FLAIR, 3D FLAIR, T1w sMRI) of cognitively healthy older people (baseline $N = 231$, age range 64 – 87 years) with a relatively low WMH load.

As a reference for the segmentation accuracy of the algorithms, completely manually segmented gold standards were used separately for each MR image modality. To validate the algorithms, we correlated the automatically extracted WMH volumes with the Fazekas scores, chronological age, and between the time points. In addition, we analyzed conspicuous percentage WMH volume increases and decreases in the longitudinal data between two measurement points to verify the segmentation reliability of the algorithms.

All algorithms showed a moderate correlation with chronological age except BIANCA with the 2D FLAIR image input only showed a weak correlation. FreeSurfer fundamentally underestimated the WMH volume in comparison with the gold standard as well as with the other algorithms, and cannot be considered as an accurate substitute for manual segmentation, as it also scored the lowest value in the DSC compared to the other algorithms. However, its WMH volumes correlated strongly with the Fazekas scores and showed no conspicuous WMH volume increases and decreases between measurement points in the longitudinal data. BIANCA performed well with respect to the accuracy metrics – especially the DSC, H95, and DER. However, the correlations of the WMH volumes with the Fazekas scores compared to the other algorithms were weaker. Further, we identified a significant

amount of outlier WMH volumes in the within-person change trajectories with BIANCA. UBO Detector's WMH volumes achieved the best result in terms of cost-benefit ratio in our study. Although there is room for optimization with respect to segmentation accuracy (especially for the metrics DSC, H95 and DER), it achieved the highest correlations with the Fazekas scores and the highest ICCs. Its performance was high for both input modalities, although it relies on a built-in single-modality training dataset, and it showed reliable WMH volume estimations across measurement points.

Keywords

White matter hyperintensities
Automated segmentation
Brain 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 (Caligiuri et al., 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 ClassificationAlgorithm) (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 **Table 1.**

In this study, we aim to provide complementary information on the performance of the three WMH extraction algorithms and the effects of the three different modalities (T1w, 3D FLAIR, 2D FLAIR). We applied the algorithms to longitudinal MRI data of cognitively healthy older people (baseline $N = 231$, study interval = four years). First, we used three fully manually segmented gold standards (based on 16 images per modality) to cross-sectionally compare the segmentation accuracy of the algorithms using different metrics, such as the Dice Similarity Coefficient (DSC), the Outline Error Rate (OER), the Detection Error Rate (DER), and the modified Hausdorff distance for the 95th percentile (H95). Second, we statistically compared the WMH volumes estimated by the algorithms as well as the correlation of those WMH volumes with the Fazekas scores in a subset of 162 subjects containing all three image modalities. Third, we used the full longitudinal dataset to validate the three algorithms by examining the correlations of the outputted WMH volumes (a) with the Fazekas score ratings for clinical validation and (b) with chronological age. In addition, we run correlations of the outputted WMH volumes between the time points of data acquisition. Based on the results of the first three analysis steps, we performed additional exploratory analyses to study the variability of the WMH segmentation between the measurement points more closely.

Table 1

Table of the articles which validated the methods in this study, listed according to Dice Similarity Coefficient (DSC), sensitivity, and false positive ratio (FPR). This table does not claim to be complete.

Article	Algorithm	Sample	Subjects	WMH load of the subjects	GS	Inter-rater agreement for the GS	Type of algorithm validation	DSC between GS and algorithm	Sensitivity between GS and algorithm	FPR between GS and algorithm	Additional results between GS and algorithm	Population based study
Ajilore et al. (2014)	**FreeSurfer**	*N = 126* (*n* = 53 *n* = 73)	LLD[a] HC[a]	*	manually (*N* = 20, LLD)	–	cross-sectionally	–	–	–	$r = 0.91$, $p < 10^{-7}$	–
Olsson et al. (2013)	**FreeSurfer**	*N* = 152	MCI[a] (incl. dementia)	*	not totally manually, with MRIcron (*N* = 27)	yes (*N* =2)***	cross-sectionally	–	–		2D FLAIR (GS) vs T1w (FS) Spearman's Rho = 0.65 ICC = 0.51 Kendall's tau = 0.48**	Gothenburg MCI study
Samaille et al. (2012)	**FreeSurfer**	1) *N* = 24 2) *N* = 43	1) MCI[a], 2) CADASIL[a]	1) * 2) very high	1) manually 2) FLAIR images as base using BioClinica SAS	–	cross-sectionally	0.40	–	–	ICC = 0.52	–
Smith et al. (2011)	**FreeSurfer**	*N* = 147 (*n* = 40, *n* = 96, *n* = 11)	normal cognition MCI[a] AD[a]	*	manually (*N* = 10)	_	cross-sectionally	–	–	–	Intraclass Correlation Coefficient = 0.91	Community based study
Jiang et al. (2018) (developers)	**UBO Detector**	1) *N* = 400 2) *N* = 539 at baseline	1) + 2) incl. stroke, TIA[a], AF[a], depression, dementia (not at baseline)	low – high*	manually (*N* = 40)	–	1) cross-sectionally 2) longitudinally	0.848 (overall)	0.913 (overall)	0.026 (overall)	overall Specificity = 0.989 overall Accuracy = 0.989 overall OER = 0.224 overall DER = 0.039 **	1) Older Australian Twins Study (OATS) 2) Sydney Memory and Ageing Study (Sydney MAS)
Griffanti et al. (2016) (developers)	**BIANCA**	1) *N* = 85 2) *N* = 474	1) AD[a], MCI[a], subjective CI[a], HC[a] 2) (neurodegenerative cohort = NDGEN) non-disabling stroke, TIA (vascular cohort = OXVASC)	medium*	manually 1. *N* = 21 2. *N* = 109	_	cross-sectionally	1) 0.76 2) 0.52	_	1) 0.22 2) 0.46	1) ICC = 0.990, DER[a] = 0.03, OER[a] = 0.46** 2) ICC = 0.919, DER[a] = 0.19, OER[a] = 0.76**	1) Oxford Project to Investigate Memory and Ageing (OPTIMA) 2) Oxford Vascular Study (OXVASC)
Ling et al.(2018)	**BIANCA**	1) *N* = 90 (2D FLAIR 2) *N* = 66/90 (3D FLAIR)	CADASIL[a]	very high	semi-automated (*N* = 20)	yes (*N* =2)***	cross-sectionally	1) median 0.79 2) median 0.76	_	1) 0.23 2) 0.20	1) ICC = 0.81** 2) ICC = 0.78**	CADASIL study
Sundaresan et al. (2018) (developers)	**LOCATE**	1) *N* = 21 2) *N* = 18 3) *N* = 15 4) *N* = 19 5) *N* = 60	1) + 2) same as in Griffanti et al. (2018) 3) CADASIL[a] 4) HC 5) from the WMH segmentation study MWSC, see Kuijf et al. (2019)	low – very high	1) + 2) see Griffanti et al. (2018) 3) no GS 4) no GS 5) manually (Kuijf et al., 2019) For 3) + 4) they used BIANCA trained with 2) as a reference – no manually traced GSs	5) yes (N =2)***	cross-sectionally	1) 0.77 2) 0.75 3) 0.79 4) – 5) range = 0.63 – 0.73	1) 0.03 (increase to global threshold of BIANCA) 2) 0.10 (increase) 3) 0.48 (increase) 4) – 1), 2), 5) for PVWMH and DWMH **	1) 0.001(increase) 2) 0.002 (increase) 3) 0.00 4) – 5) small increase**	–	–

Notes: GS = gold standard; Tp = Time point; *N* = Number of subjects; ICC = Interclass Correlation Coefficient, OER = Outline Error Rate.
[a] Subjects' clinical status: MDD = major depression disorder, LLD = late-life depression, HC = healthy controls, MCI = amnestic mild cognitive impairment; CADASIL = Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy, AD = Alzheimer's disease, TIA = transient ischaemic attack, AF = arterial fibrillation, CI = cognitive impairment.* no clear information; ** for more results/information see article; *** The number of *N* takes into account the number of operators included in the calculation for the inter-operator reliability.
If multiple results are given, those based on the highest DSC are given.

2 Material and 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, right-handedness, fluent German language proficiency, a score of ≥ 26 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 $N = 166$ datasets (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. T1-weighted (T1w) and 2D-FLAIR structural images were part of the standard MRI battery and are therefore available for the most part. 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 MR modality (T1w, 2D FLAIR, 3D FLAIR) and data acquisition time points. While the 2D FLAIR + T1w and 3D FLAIR + Tw1 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.

2.3 Subsets

In this work we used three subsets to validate the algorithms. The datasets differ in MRI sequence and number of images.

2.3.1.1 Subset 1 (input FreeSurfer)

Subset 1 included 800 sessions, containing one or two T1w images. For details on the number of subjects per time point, see **Table 2.**

2.3.1.2 Subset 2 (input UBO Detector and BIANCA)

Subset 2 consisted of 762 MR sessions, including a 2D FLAIR, and one or two T1w images. For details on the number of subjects per time point, see **Table 2.**

2.3.1.3 Subset 3 (input UBO Detector and BIANCA)

Subset 3 comprised 166 MR sessions, including a 3D FLAIR, and one or two T1w images. For details on the number of subjects per time point, see **Table 2.**

2.3.1.4 Subset *«n162»*

This subset «*n162*» (age: *72.3*; range = 65.1 – 83.9; F:M = 66:96) represents a subset of the original LHAB dataset and was built by including only sessions (*n* = 162) containing all three imaging modalities (T1w, 3D FLAIR, 2D FLAIR).

Table 2

Number (N) of sessions per time point (Tp), and in total per modality.

	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

2.4 Validation metrics

We use different 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 gold

223 standards amongst each other and (b) the algorithm outputs with the gold standards. The equations
224 can be found in **Table 3**.
225

226 **2.4.1 Overlap agreement**

227 If voxels are correctly classified in a binary segmentation, they can be true positives (TPs) or true
228 negatives (TNs). In contrast, when there is a discrepancy between the gold standard and the
229 algorithm, then the voxels are false positives (FPs) or false negatives (FNs). In case of a low WMH
230 load, as in this work, the number of TPs is much smaller than the number of TNs, which can affect
231 the accuracy measures. The FPR, also known as sensitivity, was calculated and mentioned. The
232 specificity (also known as recall) was not provided since it is equal to 1 - FPR.
233 The Dice Similarity Coefficient (DSC) provides information about the overlap agreement between
234 two segmentations (operators or automated segmentation methods) and it is perhaps the most
235 established metric in evaluating the accuracy of WMH segmentation methods. However, since the
236 DSC depends on the lesion load (the higher the lesion load, the higher the DSC), it is difficult to
237 evaluate operators or automated segmentation methods against each other if assessed on different sets
238 of scans with different lesion loads (Wack et al., 2012).
239 Extending the DSC, the Outline Error Rate (OER) and Detection Error Rate (DER) are independent
240 of lesion burden (Wack et al., 2012). In these metrics, the sum of FP and FN voxels is split,
241 depending on whether an intersection occurred or not. The sum is then divided by the mean total area
242 (MTA) of the two operators to obtain a ratio with the DER as a metric of errors without intersection,
243 and OER as a metric of errors where an intersection is found.
244

245 **2.4.2 Resemblance agreement**

246 The Hausdorff distance is a shape comparison method and can be used to evaluate the degree of
247 resemblance of two images (Beauchemin, Thomson, & Edwards, 1998; Huttenlocher, Klanderman,
248 & 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).

2.4.3 Volumetric agreement

For the volumetric agreement, an Interclass Correlation Coefficient (ICC) can be calculated additionally to identify the reliability. 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[b])	$\frac{V_A + V_B - 2 * V_{A \cap B}}{MTA}$	$\frac{2 * (FP + FN)}{\mathrm{FP} + \mathrm{FN} + 2 * \mathrm{TP}}$
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).
[b] MTA = mean total area, area of rater A and area of rater B divided by 2 (Wack et al., 2012).

2.5 Gold standard measures

To provide a valid baseline for the evaluation of the WMH segmentation accuracy of the different algorithms, WMH were (a) manually segmented in a subsample of images and (b) visually rated using the Fazekas scale (Fazekas et al., 1987). We refer to these segmentations and ratings as gold standard measures given that the manually segmented WMH represent strong proxies of the WMH load «ground truth».

2.5.1 Manual segmentation

Selection of the subsample: WMH of a subsample of all participants of the LHAB database were manually segmented on three different MRI modalities (T1w, 2D FLAIR, 3D FLAIR), resulting in three binary masks with the values *0* for background and *1* for WMH. Only participants, for which T1w, 2D FLAIR and 3D FLAIR images were available across all data acquisition time points, were considered for the manual segmentation subsample. Of those, datasets were chosen based on the median values of the Fazekas scores (see Validation of gold standard) to adequately represent the whole sample. The remaining datasets were filtered according to the MRI images available and visually inspected to find the most representative images.

Further, as an algorithm has to deal with special conditions, such as milky regions and silent lacunes (SL), five additional datasets comprising such special conditions were chosen to be included in the manual segmentation subsample. This procedure led to a selection of sixteen images.

Segmentation of FLAIR images: Three operators (O1, O2, O3) segmented the WMH 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. Sixteen 3D FLAIR images and ten 2D FLAIR images were manually segmented. The segmentations were carried out independently resulting in three different masks per operator. Six additional 2D FLAIR images were segmented by one operator (O2) obtaining the required training dataset for BIANCA (for more details see Validation of the gold standard), and resulting in an equal number of images per modality. All

operators were trained and supported for several months by *S.K.*, a neuroradiology professor with over 30 years of experience in diagnosing cerebral MRI images. All MRI sequences were fully manually segmented in all three planes (sagittal, coronal, axial) using FSLeyes (McCarthy, 2018). The software allows simultaneous viewing during segmentation in the coronal, sagittal and axial planes.

Segmentation of T1w images: Two operators (O1, O2) split the 16 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.*.

Mean mask: The three manually segmented masks of the same MR image were displayed as overlays in FSLeyes in order to evaluate mask agreement across the 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 ("**insert Supplementary Figure 1 here**"). Each mask overlay was then revised by consensus in the presence of all operators (O1, O2, O3), using the same MacBook Pro 13 inch with full brightness intensity 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.

Validation of the gold standard: The mean dice similarity coefficient (DSC) between all three operators for the 3D (n = 16) and 2D FLAIR images (n = 10) was 0.73 and 0.67, respectively. The 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, DSC is lower for images with lower WMH load (Wack et al., 2012). In this case, a DSC above 0.5 is still considered as a very good agreement (Dadar et al., 2017). Our average result for both modalities for the medium WMH load was higher than 0.7, and for the low WMH load higher than 0.6, which can be considered as an excellent agreement. The reliability of the volumetric agreement between the segmentations of the 3D and 2D FLAIR images, as indicated by the ICC as 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 "**insert Supplementary Table 1 here**".

In preparation for the optimization phase of the UBO Detector (Jiang et al., 2018), and for the mandatory training dataset for BIANCA (Griffanti et al., 2016), six additional 2D FLAIR images (the remaining six to obtain a subsample size of n = 16 images also for the 2D FLAIR sequence) were manually segmented. Because of the inter-operator reliabilities in the first ten 2D FLAIR segmentations, 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 4 shows an overview of the mean WMH volumes resulting from the manual segmentations based on with the 16 images per sequence. No significant mean WMH volume differences were revealed between the three different MR 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 WMH volume differences between the 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 "**insert Supplementary Figure 2 here**").

Table 4

Mean WMH volume in cm^3 of manually segmented gold standard (GS) with the same 16 images each modality.

	T1w GS	**3D FLAIR GS**	**2D FLAIR GS**
Total WMH volume	7.781 cm^3	8.311 cm^3	9.032 cm^3

Notes: No significant differences were found between the modalities (Kruskal-Wallis, Post-Hoc Dunn-Bonferroni with Holm correction).

2.6 Visual ratings

The Fazekas scale is a widely used visual rating scale that provides information about the location of WMH lesions (periventricular WMH (PVWMH) and deep WMH (DWMH)) as well as the severity of the WMH lesions. It ranges from 0 to 3 for both domains, 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 and 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. If a 3D FLAIR image was available it was used for the rating, if none was available the 2D FLAIR was taken, in a few cases the T1w image had to be used. Again, the ratings were carried out independently by the three operators validated for the further procedure with the following statistical indicators.

Validation of the Fazekas scale: The inter-operator agreements 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 data acquisition time point. Furthermore, 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). Finally, Spearman's rho was calculated based on subsets 1–3 to investigate the correlation of the extracted WMH volumes of the different algorithms and the ordinally scaled Fazekas scores. Therefore, the median score was calculated for each participant for each time point, split into total WMH, PVWMH and DWMH. For presentation reasons the WMH volumes were log-transformed (**see Figure 1**, panel A; **Figure 2**, panel A and B; **Figure 3**, panel A and B). The correlation between WMH volume and Fazekas score for FreeSurfer could only be calculated for the total WMH volume because its output does not discriminate between PVWMH and DWMH.

The median Fazekas scores for all three operators were: total WMH = 3; PVWMH = 2; DWMH = 1. For more descriptive details see ("**insert Supplementary Table 3 here**").

The mean inter-operator concordances across all time points over all three operators were strong (Moslem et al., 2019) according to Kendall's coefficient of concordance for total WMH ($W = 0.864$, $p < 0.001$), for PVWMH ($W = 0.828$, $p < 0.001$), and for DWMH ($W = 0.842$, $p < 0.001$). The mean inter-operator reliabilities according to weighted Cohen's kappa (Cohen, 1968) between the three operators over the four time points were substantial to almost perfect (Landis & Koch, 1977) (see "**insert Supplementary Table 2 here**").

2.7 Automated WMH segmentation

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 41 manually segmented training data (Fischl et al., 2002) including hypointensities in the white (WMH) and grey matter (non-WMH). The T1w images (subset 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 (Jiang et al., 2018) uses T1w and FLAIR images as input. We conducted the analysis once with subset 2, and once with subset 3. The probability of WMH is calculated by applying a classification model trained by 10 manually segmented 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. Segmentation of the T1w images failed in the processing of five images, whereupon these images were excluded from the following procedures. After visualizing WMH volumes once over time and once over chronological age, a massive slope was noticed in one participant. Visual inspection of the data for this participant uncovered a segmentation error (the eyeballs have been marked as WMH), thus, this time point was excluded from further analysis leading to a total number of data points of $N = 756$.

395 #### 2.7.2.1 Optimizing the k value and the threshold

396 To determine which settings are best suited for our datasets, we have evaluated the performance of
397 four different settings proposed by (Jiang et al., 2018), using the leave-one out cross-validation
398 method. Since we had manually segmented the WMH for both, 3D and 2D FLAIR images, we
399 examined the performance of UBO Detector (Jiang et al., 2018) separately for each sequence. To do
400 so, we calculated the validation metrics separately for the different settings, and checked which
401 adjustments achieved the most optimal values ("**insert Supplementary Table 4 here**"). For 3D
402 FLAIR images, UBO Detector worked most accurately with a threshold of 0.7 and a KNN of $k = 5$.
403 For the 2D FLAIR images, the best performance was achieved with a threshold of 0.9 and a KNN of
404 $k = 3$. For the subsequent calculations we used these optimized settings.
405

406 ### 2.7.3 BIANCA

407 For BIANCA (Griffanti et al., 2016), a training dataset is mandatory. As an output BIANCA
408 generates a probabilistic map of WMH for total WMH, PVWMH and DWMH. We performed the
409 calculations once with subset 2 and once with subset 3. The 16 manually segmented gold standard
410 masks derived from 3D and 2D FLAIR images were used as training dataset. For defining the
411 PVWMH we adopted the 10 mm distance rule from the ventricles (DeCarli et al., 2005), which was
412 also suggested by Griffanti et al. (2016). To reduce false positive voxels in the gray matter, and at the
413 same time only localize WMH in the white matter, we applied a WM mask. For the BIANCA options
414 we chose the ones that Griffanti et al. (2016) indicated as the best in terms of DSC and cluster-level
415 false-positive ratio: MRI modality = FLAIR + T1w, spatial weighting = «1», patch = «no patch»,
416 location of training points = «noborder», number of training points = number of training points for
417 WMH = 2000 and for non-WMH = 10000. For more details on the descriptions and the options, see
418 Griffanti and colleagues (2016).

2.7.3.1 Preprocessing

The preprocessing steps applied before the BIANCA segmentation procedure were performed with a nipype pipeline (v1.4.2; (Gorgolewski et al., 2011)) as follows: Based on subject-specific template created by the anatomical workflows of fMRIprep (v1.0.5; (Esteban et al., 2019)), a WM-mask (FSL's make_bianca_mask command) and distancemap (distancemap command) were created. The distancemap was thresholded into periventricular and deep WM (cut-off = 10 mm). For each session, T1w images were bias-corrected (ANTs v2.1.0; (Tustison et al., 2010)), brought to the template space, and averaged. FLAIR images were bias-corrected, and the template-space images were brought into FLAIR space using FLIRT (Jenkinson & Smith, 2001).

2.7.3.2 Threshold optimization

To select the best threshold for the probabilistic output of BIANCA we first used the leave-one out cross-validation method to calculate the different validation metrics separately for the 3D FLAIR and 2D FLAIR gold standard images. The global threshold values 0.90, 0.95, 0.99. were applied, with the threshold of 0.99 for both FLAIR sequences proving to be the best fitting. For a more detailed overview see ("**insert Supplementary Table 5 here**").

2.7.3.3 LOCally Adaptive Threshold Estimation (LOCATE)

LOCATE is a method that, in contrast to a global threshold, determines spatially adaptive thresholds in different regions in the probability map. This allows to overcome the influence of spatial heterogeneity of lesion probabilities due to changes in lesion contrast, load and distribution on the final threshold map of WMH. As input, LOCATE uses the lesion probability map at subject level obtained from a WMH detection algorithm. For more details on the descriptions see Sundaresan et al. (2018). Before applying LOCATE, we normalized the FLAIR images' values within the brain masks to a range of *0* to *1* to avoid different ranges of intensities between the images.

2.8 Analysis plan

In a first step, we evaluated segmentation accuracy of the three algorithms by voxel wise comparing their segmentation outputs to the manual segmentations (i.e., gold standards). In this step we used the 16 brain images, which were manually segmented to build the gold standard, and classified each voxel of a given WMH mask outputted by the algorithms as TP, TN, FP or FN. Based on these numbers we calculated the DSC for a given mask as compared to its specific gold standard (i.e., UBO Detector segmentation based on 3D FLAIR images vs. manual segmentations based on 3D FLAIR images). The DCS (see 2.5.1 Manual segmentation) was calculated across all 16 images and for different levels of WMH load (i.e. low: < 5cm^3; medium: 5–15 cm^3; high: > 15 cm^3). Besides the DCS, we calculated the following metrics: Sensitivity, H95, FPR, DER and OER. In addition, we statistically compared the automatically extracted WMH volumes with the gold standard WMH volumes by means of a Wilcoxon rank-sum test and calculated the ICC to quantify volumetric agreement.

In preparation of a second step, we compared the two thresholding options of BIANCA (global thresholding vs. LOCATE, see 2.7.3.3 LOCally Adaptive Threshold Estimation (LOCATE) to investigate which option had the better segmentation quality. For this we used the leave-one out cross-validation method with the 16 brain images, which were manually segmented to build the gold standard, and calculated the same metrics as in the first step. Because LOCATE did not perform better than the global thresholding with 0.99 we used the latter analysis pipeline for further analyses.

In step two, the accuracy of the three algorithms in relation to the gold standard – measured with the mentioned metrics – was compared using the 16 images per modality. With the subset *«n162»* the WMH volumes of the algorithms, and further the correlations of the Fazekas scores with the WMH volumes could be compared with a larger data set. The comparisons of the algorithms with the accuracy metrics (n = 16) and the comparisons of WMH volumes of the algorithms (n = 162) were compared using the Kruskal-Wallis test and analyzed post-hoc with the Dunn-Bonferroni (Holm

correction). Further we compared these outputted WMH volumes with the effect sizes according to Cohen's *d* (Cohen, 1992). With the Spearman's rank correlation, we correlated the Fazekas scores with the WMH volumes of the algorithms (n = 162), and compared them by using the effect sizes according to Cohen's q. The ICC's of the algorithms were interpreted with a degree of reliability according to Cicchetti (1994). The Wilcoxon-rank-sum test was used to determine whether the WMH volumes of the algorithms differ from the WMH volumes of the gold standards (n = 16) (effect sizes according to Cohen's *d*).

In a third step, we validated the three algorithms by examining the correlations of the outputted WMH volumes (a) with the Fazekas scores ratings for clinical validation and (b) with chronological age, which is known to be positively related to WMH volume (D. M. J. van den Heuvel et al., 2006). In addition, we ran correlations of the outputted WMH volumes between the four time points of data acquisition. For this third step we used the predefined subsets 1, 2 and 3. For the validation of FreeSurfer subset 1 (T1w only) was used, while the validation of UBO Detector and BIANCA relied on subset 2 (2D FLAIR + T1w) and 3 (3D FLAIR + T1w).

Analysis step 4 was exploratory and based on our observation of strong fluctuations of the WMH volumes extracted with BIANCA between two time points. To further investigate the variability of WMH volumes in within-person change trajectories for the algorithms and modalities, we determined the percentage and number of «conspicuous intervals between two measurement points» and also of «subjects with conspicuous longitudinal data» based on the mean percentages of WMH volume increases (mean + 1SD) and decreases (mean - 1SD), separately for the different intervals (1-, 2-, 3-, 4-year intervals) to calculate a «range of tolerance» and exclude conspicuous data points. The conspicuous 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 «conspicuous intervals between two measurement points» exceeded the number of «subjects with conspicuous longitudinal data», this indicated peaks or even several conspicuous data points in a single person – and would be an

indication of the zigzag pattern over time. For a more detailed description see 3.4.1 Longitudinal comparisons.

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 comparisons of the three WMH extraction algorithms with the respective gold standard (n = 16). In the second part the validations of the three algorithms with the subsets (see **Table 2**) and the *«n162»* comparison subset are presented, thus exploiting as much of the LHAB data as possible.

3.1 Comparisons of the algorithms

In this first subsection we will compare the results of the algorithms with the different modalities. First, the 16 gold standards per modality were used to compare the already mentioned metrics (DSC, H95 etc.), the ICCs, and the WMH volumes of the algorithms with those of the gold standards. Furthermore, the «n162» subset was used to compare the WMH volumes provided by the algorithms but also to interpret the correlations of the Fazekas scores with the WMH volumes of the different algorithms.

3.1.1 Confusion matrix

To better understand the differences of the outcomes, the values can be discussed in terms of a confusion matrix based on the gold standards by calculating the mean TPs, mean TNs, mean FPs, and mean FNs in the number of voxels and in percent (see **Table 5**).

Table 5

Confusion Matrix for the three methods, with the total number (N) of voxels per modality and further the true mean true positives, true negatives, false positives and false negatives in percent and the corresponding number of voxels.

Method Modality			Mean TP	Mean TN	Mean FP	Mean FN
FreeSurfer T1w	**Total *N* voxel**	16777216				
	***N* voxel**		2660.8	16768'659.2	775.4	5'120.5
	Percent		0.016	99.949	0.005	0.031
UBO 3D FLAIR	**Total *N* voxel**	21364736				
	***N* voxel**		7673.2	21345497.8	3552.9	8011.8
	Percent		0.036	99.910	0.017	0.038
BIANCA 3D FLAIR	**Total *N* voxel**	21364736				
	***N* voxel**		9939.1	21343431.1	5619.6	5746
	Percent		0.047	99.900	0.026	0.027
UBO 2D FLAIR	**Total *N* voxel**	10035200				
	***N* voxel**		4286.0	10024079.3	3866.2	2968.5
	Percent		0.043	99.889	0.039	0.030
BIANCA 2D FLAIR	**Total *N* voxel**	10035200				
	***N* voxel**		5937.1	10021871.6	3512.6	3878.8
	Percent		0.059	99.867	0.035	0.039

Notes: Matrix with the respective 16 gold standards.
TPs = true positives; TNs = true negatives; FPs = false positives; FNs = false negatives.

3.1.2 Dice Similarity Coefficients

As demonstrated in **Table 6**, the DSC of all algorithms is sensitive to the WMH load,

categorized according to the respective gold standard. 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 medium to high WMH exposure in terms of DSC.

Table 6

Comparison of the mean Dice Similarity Coefficient (DSC) of the different algorithms 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 16 images per modality.

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 ± SD		0.434 ± 0.111	0.501 ± 0.124	0.602 ± 0.133	0.531 ± 0.113	0.561 ± 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).

3.2 Summary of the comparisons

Table 7 provides an overview of the outcomes of the inferential statistics. On the one hand, the different metrics of the algorithms are compared, using the 16 manually segmented gold standards of each sequence. On the other hand, the *«n162»* subset was used to compare the correlation between WMH volume and Fazekas score, but also to compare the total WMH volumes across modality and algorithms.

The results of the mean DSC suggest that BIANCA performed best in terms of the degree of overlap between gold standard and algorithm output. However, a significant difference could only be seen between FreeSurfer and BIANCA. The 3D FLAIR inputs indicate a slightly better value on average, with BIANCA 3D FLAIR showing the best result. Nevertheless, the methods using FLAIR images as input 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 input, was very accurate, and on average, significantly better than UBO 3D FLAIR. Visually, there are also noticeable differences to the other algorithms, see “insert **Supplementary Figure 3** here”. The accuracy of BIANCA with both modalities was also reflected by the high sensitivity. However, UBO 2D FLAIR does not differ significantly from BIANCA. Only FreeSurfer performed significantly worse than BIANCA and UBO 2D FLAIR. In contrast, FreeSurfer on average showed a significantly better FPR than all others. BIANCA 3D FLAIR achieved on average a significantly better DER than UBO, while FreeSurfer performed significantly worse regarding the OER compared to all other results. FreeSurfer was the only algorithm that on average significantly underestimated the WMH volumes compared to the gold standard. Moreover, the mean WMH volume was significantly different from all outputs of the other algorithms and from the *«n162»* subset (both with a large effect size). Furthermore, significant mean WMH volume differences between 2D FLAIR and 3D FLAIR inputs were found, with 2D FLAIR inputs tendentially leading to higher WMH volume estimations. In terms of the ICC(3,1), FreeSurfer achieved a fair, BIANCA 3D FLAIR a good, and BIANCA 2D FLAIR, UBO 3D and 2D FLAIR an excellent degree of reliability with UBO 2D FLAIR showing the highest ICC. BIANCA ranks worst in the correlation between WMH volume and the Fazekas scores using the *«n162»* subset. Within BIANCA, the 2D FLAIR sequences performed worse than the 3D FLAIR sequences with a moderate effect size. A large effect size was shown between BIANCA 2D FLAIR and FreeSurfer, UBO 2D FLAIR, as well as UBO 3D. UBO Detector showed the highest correlation of all algorithms between the total WMH volume and the Fazekas scores.

Table 7

Summary of comparison for FreeSurfer, UBO Detector and BIANCA compared to the gold standard for Dice Similarity Coefficient (DSC), Hausdorff distance for the 95th percentile (H95), sensitivity, false positive ratio (FPR), Detection Error Rate (DER), Outline Error Rate (OER), WMH volume of the gold standards versus WMH volume of the algorithms, and Interclass Correlation Coefficient (ICC). Comparison with the «n162» subset: WMH volumes of the different algorithms, and correlation between WMH volume and Fazekas score.

	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] (*n* = 16)	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] (*n* = 16)	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] (*n* = 16)	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] (*n* = 16)	.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] (*n* = 16)	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] (*n* = 16)	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
Volume (cm^3)[a] (*n* = 162)	3.369	2.758 – 3.980	6.630	5.471 – 7.790	11.373	9.687 – 13.060	7.047	6.138 – 7.957	14.086	12.098 – 16.073	< 0.001	FreeSurfer (d > 1.0)* < all others; BIANCA 2D (d = 0.98)*, UBO 2D (d = 0.78)* > BIANCA 3D, UBO 3D (d = 0.28)*
	coeff	***p*-value**	**coeff**	***p*-value**	**coeff**	***p*-value**	**coeff**	***p*-value**	**coeff**	***p*-value**	**Interpretation**	
GS vol. vs. Alg. Vol.[b] (n = 16)	*W* = 43	< 0.001	*W* = 85	0.11	*W* = 137	*0.752*	*W* = 133	0.862	*W* = 123	0.867	large effect size (d > 1.0)* between GS vol. and FreeSurfer vol.*	
ICC(3,1) (n = 16)	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**	
Fazekas score Spearman's rho[c] (*n* = 162)	0.711	< 0.001	0.802	< 0.001	0.799	< 0.001	0.582	< 0.001	0.347	< 0.001	Small effect between FreeSurfer and UBO 2D (*q* = 0.207), UBO 3D (*q* = 0.215), BIANCA 3D (*q* = 0.224), moderate effect between BIANCA 3D and BIANCA 2D (*q* = 0.303), UBO 2D (*q* = 0.430), UBO 3D (*q* = 0.439), large effect between BIANCA 2D and FreeSurfer (*q* = 0.527), UBO 2D (*q* = 0.734), UBO 3D (*q* = 0.742) ***	

Notes: [a] Comparison by Kruskal-Wallis. [b] Comparison by Wilcoxon-rank-sum-test. [c] Spearman's rho = 0.1 – 0.3 = weak, 0.4 – 0.6 – 0.7 = moderate, 0.7 – 0.9 = strong, – 1 = perfect (Dancey & Reidy, 2017)
* Cohen's d = 0.5 – 0.7 = moderate, 0.8 – ≥ 1.0 = large effect size (Cohen, 1988); 2.0 = huge effect size (Sawilowsky, 2009)
** Between 0.40 – 059 = fair, 0.60 – 0.74 = good, 0.75 and 1.00 = excellent (Cicchetti, 1994)
*** Cohen's *q* = < 1 = no effect, 1 – 3 = small effect, 3 – 5 = moderate effect, >5 = large effect (Cohen, 1988)

3.3 Validation of the algorithms

For the validation every method was used for segmenting the WMH on all images of subset 1 (800 images), subset 2 (756 images), and subset 3 (166 images). To clinically validate the algorithms we correlated the WMH volumes with the Fazekas scores but also with chronological age, which is regarded as a good external standard (D. M. J. van den Heuvel et al., 2006). To analyze the reliability of the WMH volume over time, we correlated it between the time points (Tp1 – Tp2, Tp2 – Tp3, Tp3 – Tp5). We further compared the BIANCA output with the global threshold with the output of LOCAL using the gold standard as a reference.

3.3.1 Validation of FreeSurfer

3.3.1.1 Correlation with Fazekas scale, chronological age and between the time points

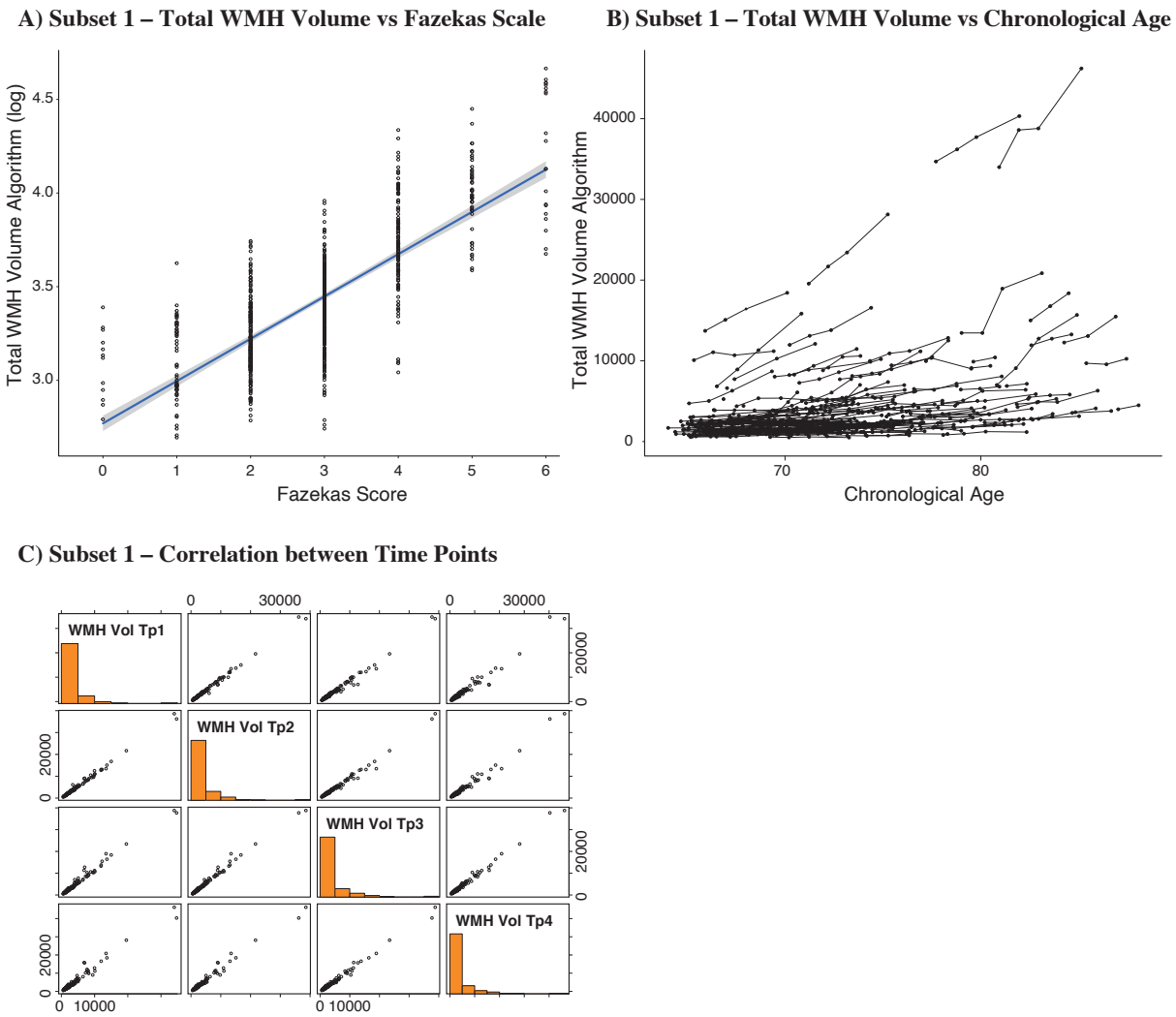
In subset 1 (800 images) a significant correlation between total WMH volume and Fazekas scores was found. The Spearman's rho correlation was strong (r_s= 0.770) according to (Dancey & Reidy, 2017). The visual distribution of the log-transformed total WMH volume with respect to the Fazekas scores is shown in **Figure 1** (panel A).

A significant correlation between chronological age and total WMH volume extracted by FreeSurfer was found in subset 1. The Spearman's rho correlation was moderate: FreeSurfer totalWMH (r_s= 0.443) (see **Figure 1**, panel B).

We also found a significant correlation for the total WMH volumes extracted by FreeSurfer between the time points. The Pearson's product-moment correlation showed an almost perfect (Dancey & Reidy, 2017) linear distribution: Tp1 to Tp2 (r = 0.997), Tp2 to Tp3 (r = 0.994), Tp3 to Tp5 (r = 0.995) (see **Figure 1**, panel C). For an overview of the results of all algorithms see **Table 9.**

Figure 1

FreeSurfer validation. *Scatter plot of the total WMH volume distribution according to Fazekas scale (A), chronological age (B), and correlation between the time points (C). The correlation values are mentioned in the main text.*



3.3.2 Validation of UBO Detector

3.3.2.1 Correlation with Fazekas scale, chronological age and between the time points

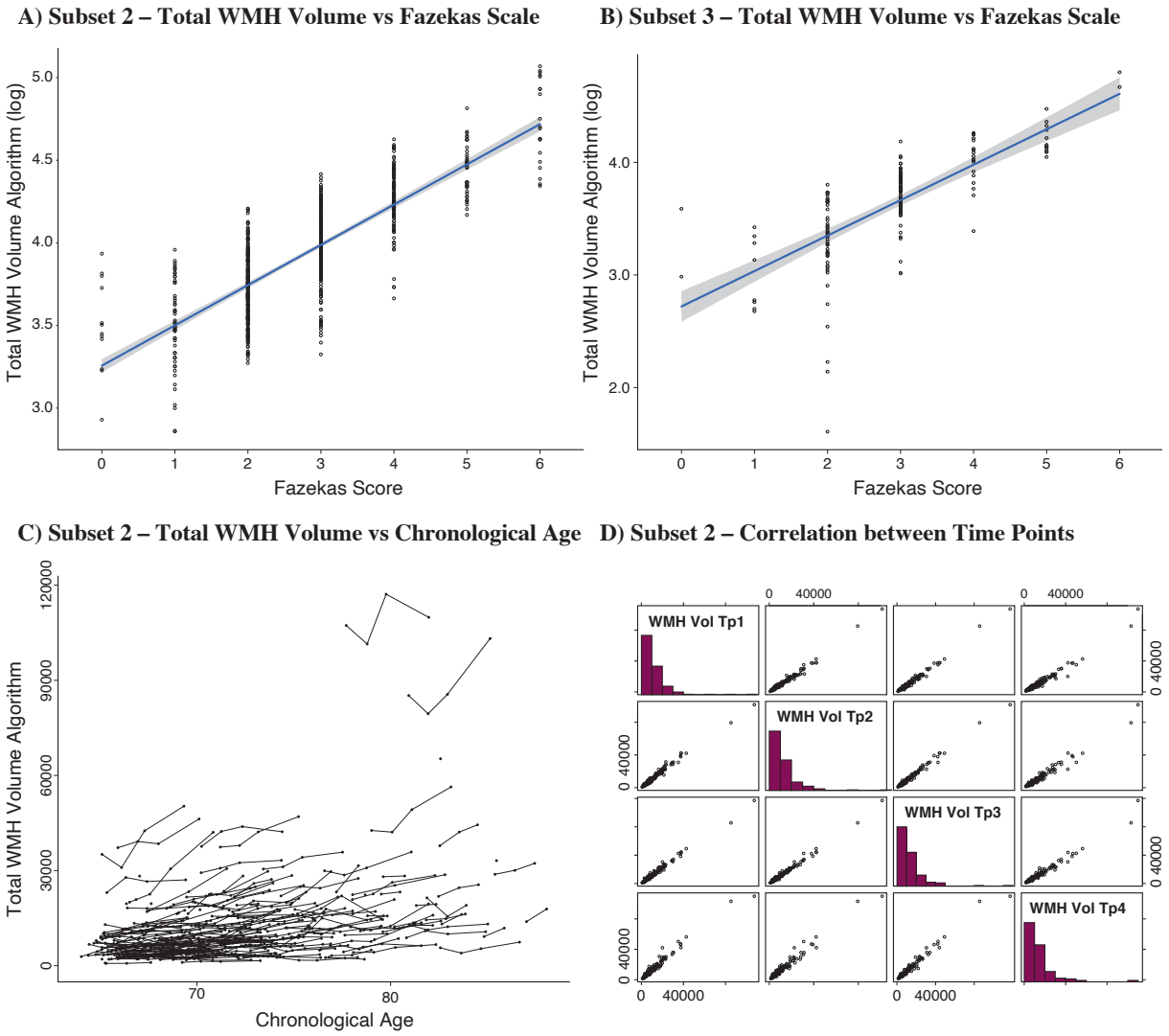
In subset 2 and 3 (756 images for subset 2, and 166 images for subset 3) significant correlations between WMH volumes and Fazekas scores were found. The Spearman's rho correlations were strong (Dancey & Reidy, 2017) for UBO Detector in subset 2: UBO totalWMH (r_s = 0.794), UBO PVWMH (r_s = 0.734), UBO DWMH (r_s = 0.615), and in subset 3: totalWMH (r_s = 0.803), UBO PVWMH (r_s = 0.770), UBO DWMH (r_s = 0.606).

The visual distribution of the log-transformed total WMH volume with respect to the Fazekas scores for subset 2 and 3 is shown in **Figure 2** (panel A and B).

A significant correlation between chronological age and total WMH volume extracted by UBO Detector was found in both subsets. The Spearman's rho correlations were moderate in subset 2: UBO totalWMH (r_s = 0.398), and in subset 3: UBO totalWMH (r_s = 0.351) (see **Figure 2**, panel C). We also found significant correlations in subsets 2 and 3 for the total WMH volumes extracted by UBO Detector between the time points. The Pearson's product-moment correlation showed an almost perfect (Dancey & Reidy, 2017) linear distribution in subset 2: Tp1 to Tp2 (r = 0.990), Tp2 to Tp3 (r = 0.992), Tp3 to Tp5 (r = 0.983) (see **Figure 2**, panel D), and in subset 3 (due to insufficient longitudinal data, only two correlations between the time points could be calculated): Tp2 to Tp3 (r = 0.998, $p < 0.001$), Tp3 to Tp5 (r = 0.970, $p < 0.001$). For an overview with the results of all algorithms see **Table 9.**

Figure 2

UBO validation. *Scatter plot of the total WMH volume distribution according to Fazekas scale (A and B), chronological age (C), and correlation between the time points (D). The correlation values are mentioned in the main text.*



3.3.3 Validation of BIANCA

3.3.3.1 Comparison of the threshold methods

In order to determine whether BIANCA with the best global threshold of 0.99 or the LOCAL method is more suitable for our data, for each method we carried out a leave-one-out cross-validation against both the 3D FLAIR and 2D FLAIR gold standards.

As shown in **Table 8** and **"insert Supplementary Figure 5 here**", on average BIANCA marks significantly less FPs compared to LOCATE. In BIANCA 3D FLAIR, the mean H95 was significantly better than in LOCATE 3D. Only the mean sensitivity was better in LOCATE than

in BIANCA. The mean WMH volume of LOCATE was significantly different from the gold standard and also significantly higher than the mean WMH volume of BIANCA. This was also reflected in the low and non-significant ICC(3,1) of LOCATE.

Based on the results in **Table 8**, the outliers of each DSC in LOCATE compared to the gold standard (DSC range: LOCATE 2D FLAIR = 0.194 – 0.687, 3D FLAIR = 0.165 – 0.705; BIANCA 2D FLAIR 0.334 – 0.734, BIANCA 3D 0.292 – 0.783), and the visual inspections, we decided to use BIANCA for the whole sample.

Table 8

Statistical comparison between BIANCA (global threshold of 0.99) and LOCATE – using the gold standard with the 16 images per modality as reference – of the different accuracy metrics, and the WMH volumes, but also the comparison between the automated methods and the manually gold standard of the WMH volumes and the ICCs.

	BIANCA 3D FLAIR[a]	**LOCATE 3D FLAIR[b]**	**WILCOXON**		**BIANCA 2D FLAIR[a]**	**LOCATE 2D FLAIR[b]**	**WILCOXON**	
			W	***p***			***W***	***p***
DSC	0.602	0.552	144	0.564	0.561	0.512	150	0.423
OER	0.635	0.706	124	0.897	0.664	0.751	106	0.423
DER	0.162	0.190	87	0.128	0.215	0.224	117	0.696
H95	6.200	9.298	54	0.004	8.443	9.314	102	0.337
FP	5619.6	19241.4	46	0.001	3512.6	10820.1	46	0.001
TP	9939.1	12761.4	99	0.287	5937.1	7605.6	97	0.254
FPR	0.0003	0.0009	46	0.001	0.0004	0.0011	46	0.001
Sensitivity	0.611	0.812	19	< 0.001	0.575	0.766	32	< 0.001
	COEFF *p*-value	**COEFF *p*-value**			**COEFF *p*-value**	**COEFF *p*-value**		
Vol. method vs Vol. GS[a]	*W* = 133 *p* = 0.867	W = 198 *p* = 0.007			*W* = 123 *p* = 0.867	W = 194 *p* = 0.012		
ICC(3,1)	0.743 < 0.001	0.206 *p* = 0.141			0.859 < 0.001	0.526 *p* < 0.05*		
Vol. in cm³	8.317	17.106			8.695	16.955		
Vol. BIANCA vs Vol. LOCATE[a]	*W* = 65 *p* = 0.017				*W* = 64 *p* = 0.015			

Notes: [a] Comparison by Wilcoxon-rank-sum-test.
DSC = mean DSC; OER = Outline Error Rate; DER = Detection Error Rate; H95 = Hausdorff distance for the 95 percentile; FP = false positives; TP = true positives; FPR = false positive rate; Vol. = WMH Volume; ICC = Interclass Correlation Coefficient

3.3.3.2 Correlation with Fazekas scale, chronological age and between the time points

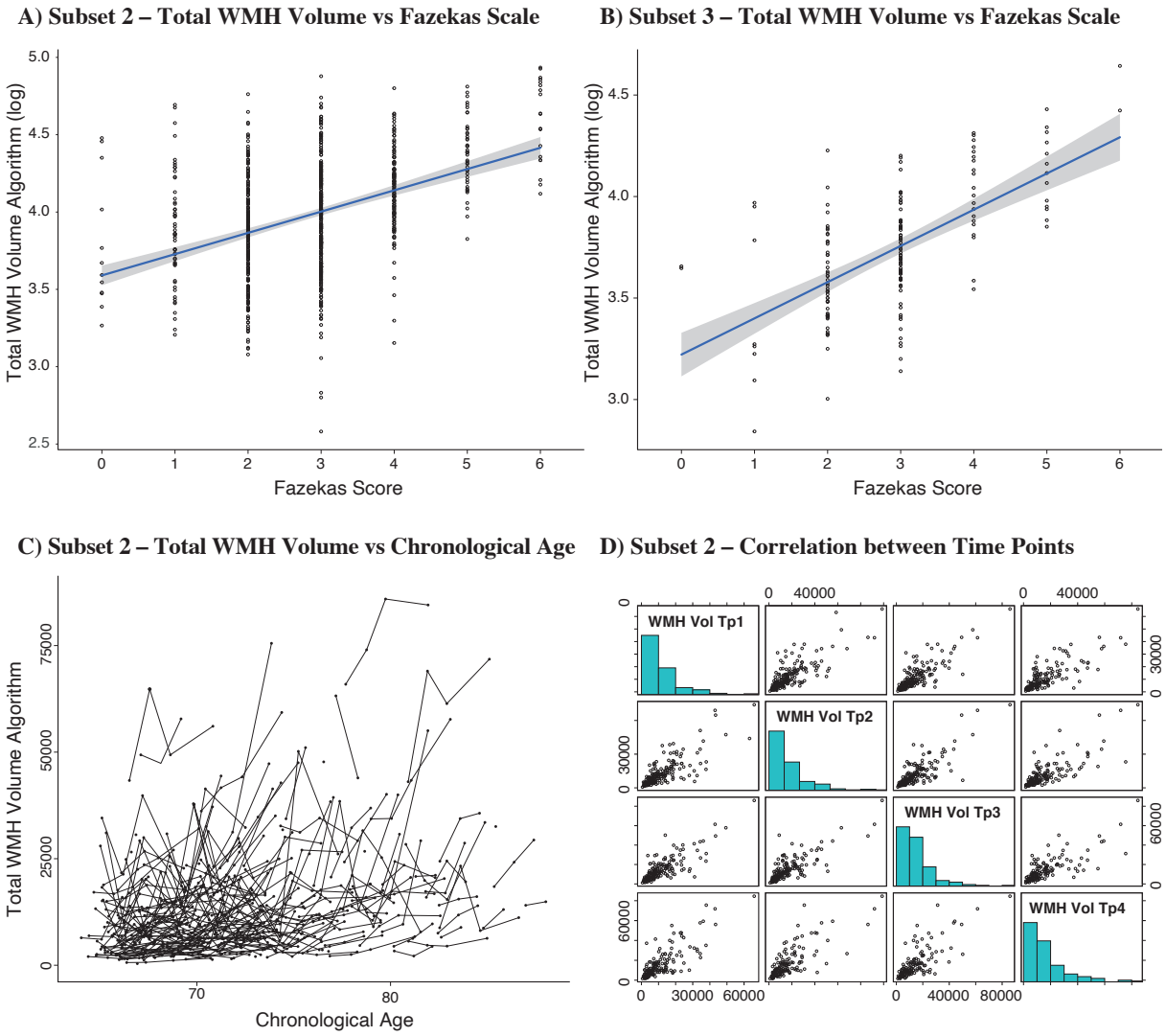
In subset 2 (762 images) and subset 3 (166 images) significant correlations between WMH volumes and Fazekas scores were found. The Spearman's rho correlations were moderate (Dancey & Reidy, 2017) for BIANCA in subset 2: BIANCA totalWMH (r_s = 0.401), BIANCA PVWMH (r_s = 0.361), BIANCA DWMH (r_s = 0.345), and in subset 3: BIANCA totalWMH (r_s = 0.577), BIANCA PVWM (r_s = 0.557), BIANCA DWMH (r_s = 0.410). The visual distribution of log-transformed total WMH volume with respect to the Fazekas scores for subset 2 and 3 is shown in **Figure 3** (panel A and B).

A significant correlation between chronological age and total WMH volume extracted by BIANCA was found in both subsets. The Spearman's rho correlation was weak in subset 2: BIANCA totalWMH (r_s = 0.301), and moderate in subset 3: BIANCA totalWMH (r_s = 0.390) (see **Figure 3**, panel C).

We found significant correlations in subset 2 and 3 for the total WMH volumes extracted by BIANCA between the time points. The Pearson's product-moment correlation showed a strong (Dancey & Reidy, 2017) linear distribution in subset 2: Tp1 to Tp2 (r = 0.802), Tp2 to Tp3 (r = 0.807), Tp3 to Tp5 (r = 0.768) (see **Figure 3**, panel D), and in subset 3 (due to insufficient longitudinal data, only two correlations between the time points could be calculated): Tp2 to Tp3 (r = 0.716, p = 0.013), Tp3 to Tp5 (r = 0.917, p < 0.001). For an overview with the results of all algorithms see **Table 9.**

712 **Figure 3**

713 ***BIANCA validation.*** *Scatter plot of the total WMH volume distribution according to Fazekas*
714 *scale (A and B), chronological age (C), and correlation between the time points (D). The*
715 *correlation values are mentioned in the main text.*



716

717

718 ## 3.4 Additional calculations to analyze the longitudinal data

719 ### 3.4.1 Longitudinal comparisons

720 Throughout the BIANCA outputs, we detected massive fluctuations in the individual change
721 trajectories over time that could not be seen in the cross-sectional data with the gold standard. An
722 example of a segmentation of a BIANCA 2D FLAIR output over four time points of one subject
723 is illustrated in "**insert Supplement Figure 4 here**" (total WMH volume of the different time
724 points: Tp1 = 5.136 cm^3, Tp2 = 4.213 cm^3, Tp3 = 20.746 cm^3, Tp5 = 7.848 cm^3). In this example

time point 3 shows a huge derivation. These fluctuations between the time points are reflected visually in a conspicuous zigzag pattern in **Figure 3,** panel C, as well as in the BIANCA output masks in the way of false positive segmentations, see "**insert Supplement Figure 4 here**"). Further, compared to the other algorithms, BIANCA shows a lower correlation between the total WMH volume and the Fazekas score as well as between the time points (see **Table 9**), which is visible in the more widely scattered dots in **Figure 3,** panel D. As depicted in **Table 9**, BIANCA's WMH volumes show lower correlations between time points, with the Fazekas scores and with chronological age compared to FreeSurfer and UBO Detector.

Table 9

Correlations between the time points (Tp1–Tp2, Tp2–Tp3, Tp3–Tp5), total WMH volume calculated by the respective algorithm correlated with the Fazekas score (totalWMH – Fazekas) and with chronological age (total WMH – Age) for subset 1 for FreeSurfer, and subset 2 for UBO Detector and BIANCA.

	FreeSurfer ***N* = 800**	**UBO 2D FLAIR** ***N* = 756**	**BIANCA 2D FLAIR** ***N* = 762**
Tp1 – Tp2[a]	r = 0.997***	r = 0.990***	r = 0.802***
Tp2 – Tp3[a]	r = 0.994***	r = 0.992***	r = 0.807***
Tp3 – Tp5[a]	r = 0.995***	r = 0.983***	r = 0.768***
totalWMH – Fazekas[b]	r_s = 0.770***	r_s = 0.794***	r_s = 0.401***
totalWMH – Age[b]	r_s = 0.443***	r_s = 0.398***	r_s = 0.301***

Notes: [a] Pearson's product-moment correlation. [b] Spearman's rho.
Pearson's prodcuct-moment correlation coefficient and Spearman's rho = 0.1 – 0.3 = weak, 0.4 – 0.6 – 0.7 = moderate, 0.7 – 0.9 = strong, – 1 = perfect (Dancey & Reidy, 2017)
*** $p < 0.001$

Since it is crucial to know whether these within-subject fluctuations are only isolated cases or a more wide-spread problem, we evaluated the amount of conspicuous fluctuations. For this purpose, we checked the entire WMH volume output files of BIANCA 3D FLAIR and examined randomly the 2D FLAIR output WMH volume files. We further checked the BIANCA segmentation masks of the values found to be conspicuous. 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 assume that the massive peaks in the within-subject trajectories of the WMH volumes extracted by BIANCA – which were mostly evident in subjects with low to medium WMH load – are driven by erroneous segmentation of the algorithm. Since our knowledge of WMH volume changes (Shi & Wardlaw, 2016)is insufficient with 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 in the entire dataset, how many comparisons between two measurement points can be made for a given subject. We differentiate between 1-year intervals (baseline – 1-year follow-up / 1-year follow-up – 2-year follow-up), 2-year-intervals (baseline – 2-year follow-up, 2-year follow-up – 4-year follow-up), 3-year-intervals (1-year follow-up – 4-year follow-up), and 4-year-intervals (baseline-4-year follow-up). For BIANCA 2D FLAIR, for example, $N = 209$ subjects provide longitudinal imaging data (at least two time points of data). These 2D FLAIR images (762 images) enable 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 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, for each subject, algorithm-by-modality combination (e.g. BIANCA 2D FLAIR) and available time interval (1-year interval), we calculated the percent WMH volume change from the prior to the subsequent measurement point. Then, for each algorithm-by-modality combination and available time interval, the mean percent change was determined as well as the standard deviation. These metrics (mean WMH volume change, SD) were averaged across the five algorithm-by-modality combinations. These general volume changes metrics were then 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 as compared to segmentation errors.

779 The upper limit of the increases (average across algorithm-by-modality + 1SD) and the lower
780 limit of the decreases (average across algorithm-by-modality - 1SD) were used to identify the
781 conspicuous cases. **Supplementary Table 9** lists, for all time intervals, the algorithm and
782 modality-specific WMH volume changes as well as the average WMH volume change over all
783 algorithms and the associated «ranges of tolerance». WMH volume increases and decreases
784 outside of these «ranges of tolerance» were considered as conspicuous. **Table 10** provides
785 information on the number of conspicuous changes per algorithm-by-modality combination in
786 comparison to the number of possible comparisons between two measurement occasions. This
787 table also contains information on the distribution of conspicuous changes in dependence of
788 WMH load and informs about the percentage of subjects, in which conspicuous changes
789 occurred. To identify whether there is a relation between the WMH load and the conspicuously
790 segmented images, we divided them into low, medium and high WMH load using the Fazekas
791 scale. In each case, the previous measurement point in the individual change trajectories over
792 time served as the basis for the classification. The categories were divided into the following
793 categories: Fazekas score 0 – 2 = low WMH load, Fazekas score 3 and 4 = medium WMH load,
794 and Fazekas score 5 and 6 = high WMH load.
795 Compared to the other algorithms, BIANCA (2D and 3D FLAIR) clearly shows the most
796 «subjects with conspicuous longitudinal data» (2D FLAIR: *N* = 109, 52.15%; 3D FLAIR: *N* = 7,
797 17.95%) and «conspicuous intervals between two measurement points» (2D FLAIR: *N* = 161,
798 30.32%; 3D FLAIR: *N* = 8, 16.67%), see **Table 10**. Furthermore, BIANCA (2D and 3D FLAIR)
799 is the only algorithm that provides more «conspicuous intervals between two measurement
800 points» than «subjects with conspicuous longitudinal data», which indicates a zigzag pattern
801 within the subjects. Considering the WMH volume increases and decreases in all intervals,
802 BIANCA shows the highest values: In the 1-year intervals, for example, the 3D FLAIR image
803 output shows mean increase of 76.82% (SD = +174.65%; N = 9) and an average decrease of -
804 30.63% (SD = - 33.06%; N = 3), the 2D FLAIR image output shows an average increase of
805 64.80% (SD = +76.92%; N = 236) and an average decrease of -26.61% (SD = -19.41%; N =

133), see **Supplementary Table 9** for comparisons with the other algorithms. Regarding the WMH load, no clear pattern could be seen within BIANCA. However, in BIANCA 2D FLAIR 55.90% of the «conspicuous intervals between two measurement points» were images with a low WMH load, and 40.99% with a medium WMH load. With the 3D FLAIR data, more images with a medium WMH load were conspicuous – but there was less data available (BIANCA: $N = 8$; UBO: $N = 2$; FreeSurfer: $N = 0$).

Table 10

Display per algorithm with the respective input modality according the WMH load of the total number (N) of output images, number of subjects with longitudinal data (at least 2 time points), number and percentage (in brackets) of subjects with conspicuous longitudinal data, number intervals between two measurement points, and number and percentage (in brackets) of conspicuous intervals between two measurement points.

Algorithm Modality / WMH load	**N of images**	**N of subjects with longitudinal data**	**N (and %) of subjects with conspicuous longitudinal data**		**N of intervals[a] between two measurement points**	**N (and %) of conspicuous 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 loads are divided into low (< 5cm^3), medium (5 – 15 cm^3), and high (> 15 cm^3).
[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 validated and compared the performance of three freely available automated methods 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 shows deficiencies particularly with respect to segmentation accuracy (i.e, DSC) and clearly underestimates the WMH volumes. We therefore argue that it cannot be considered as a valid substitution for the gold standard. BIANCA and UBO Detector show a higher segmentation accuracy compared to FreeSurfer. When using 3D FLAIR images as input, BIANCA performed significantly better than UBO Detector regarding the accuracy metrics DCS, DER and H95. However, we identified a significant amount of outlier WMH volumes in the within-person change trajectories of the BIANCA volume outputs. Exploratory analyses of these conspicuous fluctuations indicate random false positive segmentations which contribute to the erroneous volume estimations. UBO Detector – as a fully automated algorithm – has the best cost-benefit ratio in our study. Although there is room for optimization regarding segmentation accuracy, it distinguishes itself through its excellent volumetric agreement with the gold standard in both modalities (as reflected by the ICCs) and its high correlations with the Fazekas scores. In addition, it proves to be a robust estimator of WMH volumes over time.

4.1 Validating the algorithms

FreeSurfer. The total WMH volumes from FreeSurfer correlated strongly with the Fazekas scores, between the time points and moderately with chronological age. It showed no conspicuous WMH volume increases and decreases between measurement points in the longitudinal data. These results show that FreeSurfer outputs reliable data over time. However, we also observed that the WMH volume validity of FreeSurfer's output is not given due to the

849 fundamental underestimation of WMH volume compared to the corresponding T1w gold
850 standard. This can be attributed to the fact that WMH often appear isointense in T1w sequences
851 and is herefore not detected (Wardlaw et al., 2013). Furthermore, the lower contrast of the
852 DWMH compared to the PVWMH, which is due to the lower water content in the DWMH as a
853 result of the longer distance to the ventricles, might contribute to the WMH volume
854 underestimation. FreeSurfer often omitted DWMH, a finding also reported by Olsson et al.
855 (2013). In addition, our analyses showed that FreeSurfers' underestimation of the WMH volume
856 was even more pronounced in high WMH load images (see Bland-Altmann Plot (Bland &
857 Altman, 1986) in **Figure 6,** panel C). The same bias was shown by Olsson et al. (2013) when
858 comparing the volumes of the semi-manually segmented WMH (2D FLAIR) and the FreeSurfer
859 (T1w) output. The spatial overlap performance of FreeSurfer in our study is comparable to the
860 findings in the validation study reported by Samaille and colleagues (2012) with a cohort of mild
861 cognitive impairment and CADASIL patients. Smith and colleagues (2011) reported an Intraclass
862 Correlation Coefficient (ICC) of 0.91 between Freesurfer's output and the 10 manually
863 segmented images from one operator. Other accuracy metrics were not calculated. Ajilore et al.
864 (2014) reported a high correlation of r = 0.91 between the WMH volume of FreeSurfer and the
865 WMH volume of their manual segmentation in T1w and T2w images, including 20 subjects with
866 late-life major depression. However, they have not described the manual segmentation procedure
867 in detail. Nevertheless, this WMH volume underestimations between T1w and both FLAIR
868 modalities, is in line with the STandards for ReportIng Vascular changes on nEuroimaging
869 (STRIVE), stating that FLAIR images tend to be more sensitive to WMH and therefore are
870 considered more suitable for WMH detection than T1w images (Dadar et al., 2018; Wardlaw et
871 al., 2013). However, comparisons to previous studies are difficult because (a) for most studies it
872 is not indicated whether they used fully manual gold standards or a gold standard generated by a
873 semi-automated method, (b) sample sizes have been small, (c) there is very little information on
874 sensitivity, H95, FPR, DER, OER and (d) FreeSurfer has not been applied to longitudinal data.
875 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 correlated highly with the Fazekas scores and showed reliable WMH volume increases and decreases over time, FreeSurfer cannot be considered as a valid substitute for manual segmentation due to the weak outcomes in the accuracy measures (DSC, OER, ICC) and especially due to its massive WMH volume underestimation.

UBO Detector. The total WMH, PVWMH and DWMH volumes from UBO Detector (*N* = 756) correlated strongly with the Fazekas scores and moderately with age. This is in line with a recent paper published 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. Also, the results of the volumetric agreements – calculated with ICCs – were similar to our results, especially for the 2D FLAIR images. In our analysis, we found slightly higher correlations between the time points of our longitudinal dataset 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 comparison. The difference regarding the sensitivity and overlap measurements may be due to the fact that, in contrast to Jiang et al. (2018), we did not use a customized training dataset but the 2D FLAIR built-in training dataset. UBO Detector is also declared as a fully automated method but no previously segmented gold standard can be inserted in the pipeline. Hence, whether a 2D or a 3D FLAIR input is used for UBO Detector may influence the estimated WMH volume. Our analysis indicates that the WMH volume estimated by UBO Detector depends on the modality of the FLAIR input (2D vs. 3D FLAIR). Volumes extracted from 2D FLAIR images tend to be more similar to the WMH volume of the respective gold standard while volumes extracted from 3D FLAIR images tend to underestimate the WMH volume of the respective gold standard (see "**insert Supplementary Table 6 here**", and Bland-Altmann Plot in **Figure 6,** panel A and B). For several reasons UBO's longitudinal pipeline was

903 not used in our study. First, UBO Detector requires an equal number of sessions for all subjects,
904 which would have resulted in a reduction of our sample size. Secondly, it registers all sessions to
905 the first time point, an approach which has been shown to lead to biased registration (Reuter,
906 Schmansky, Rosas, & Fischl, 2012). Lastly, comparing the two pipelines, Jiang et al. (2018) did
907 not find significant differences regarding the extracted WMH volumes. Importantly, up to now,
908 we have not found any other study that validated UBO Detector and/or compared it with other
909 WMH extraction methods and no study validated UBO Detector with 3D FLAIR sequences.
910 **BIANCA.** In order to properly compare our output data from BIANCA (Griffanti et al., 2016)
911 with the results of the original study from BIANCA we additionally run BIANCA's evaluation
912 script to calculate the same metrics ("insert **Supplementary Table 7 here**"). Our overall results
913 for the different accuracy metrics correspond more to those of their vascular cohort than to those
914 of their neurodegenerative cohort. We received quite similar results for the 2D FLAIR sequences
915 with respect to the correlations between the BIANCA WMH volumes and our gold standard
916 WMH volumes. We also obtained similarly moderate correlations for WMH volume and age as
917 they did for their neurodegenerative cohort (Griffanti et al., 2016). However, we were not able to
918 replicate the high ICCs for the volumetric agreement and the high correlations between WMH
919 volume and the Fazekas scores they received in both cohorts, neither for 2D nor for 3D FLAIR.
920 We suspect that this might be due to the false positives in our BIANCA outputs. While the effect
921 of false positives was less obvious in the cross-sectional analyses, it was uncovered because of
922 massive fluctuations in the longitudinal analyses. The developers of UBO Detector compared
923 their algorithm to BIANCA based on a sample of 40 subjects (Jiang et al., 2018). They noticed
924 that BIANCA tended to overestimate the WMH in «milky» regions, whereas the sensitivity for
925 WMH detection was higher in BIANCA than in UBO, which is in line with our findings. We
926 generally used the settings which, in the original description of the BIANCA pipeline (Griffanti
927 et al., 2016), were reported to produce the best results in terms of DSC and cluster-level false
928 positive ratio (also referred as false discovery rate (FDR)). To represent our entire dataset
929 adequately, we selected our training dataset in terms of the WMH loads based on the median

values of the Fazekas scores. Ling and colleagues (2018) showed better results with a mixed WMH load training dataset than with a training dataset with only high WMH load as Griffanti and colleagues (2016) suggested in their paper.

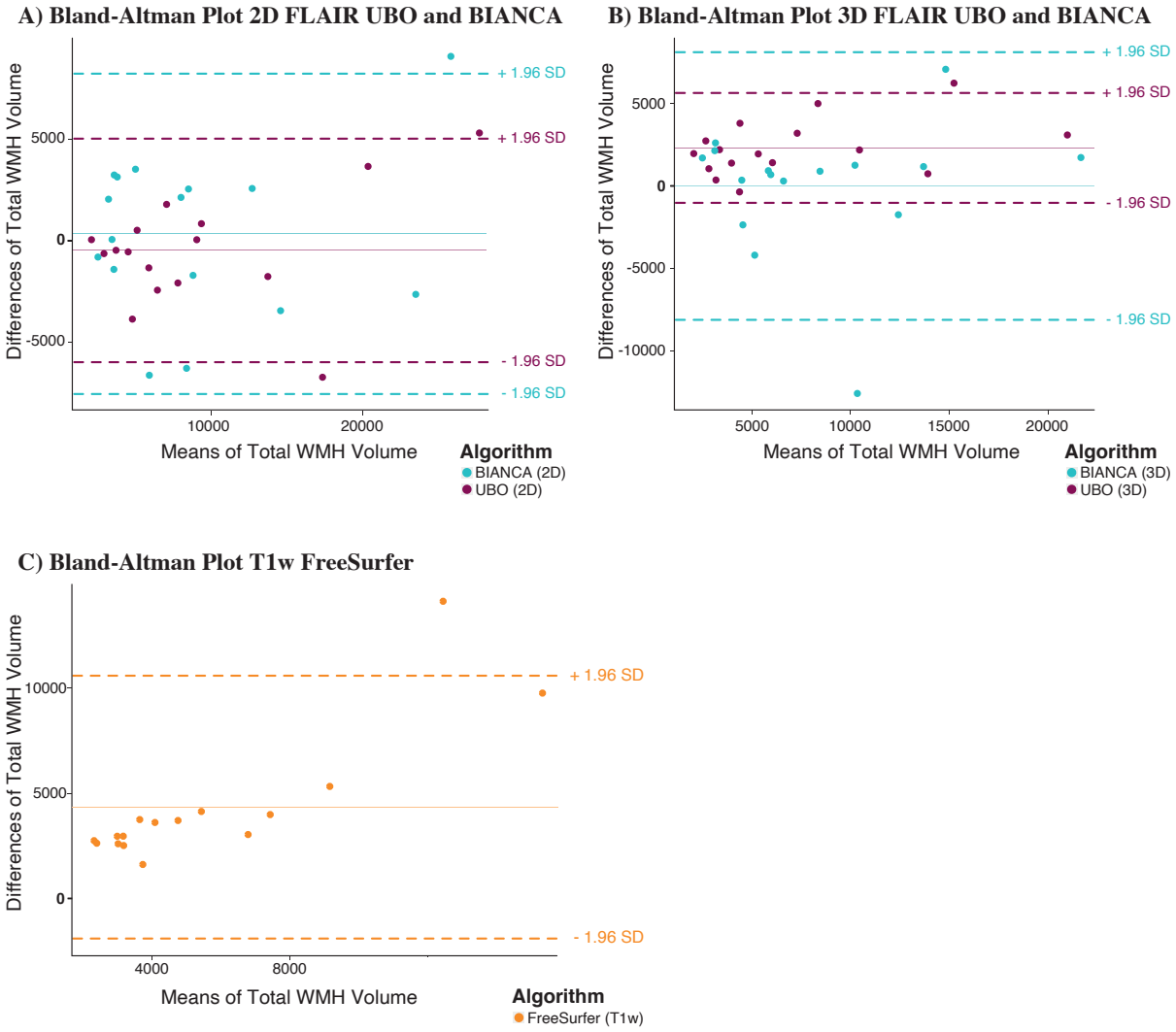
BIANCA features LOCATE as a method to determine spatially adaptive thresholds in different regions in the lesion probability map. Sundaresan et al. (2018) showed that LOCATE is beneficial when the BIANCA algorithm is trained with dataset-specific images or when the training dataset was acquired with the same sequence and the same scanner. For the group of healthy controls, they achieved similar visual outputs with LOCATE as compared to those with global thresholding. However, since no manual gold standard was available for the healthy controls in their study, a quantitative comparison between gold standard and LOCATE output was not possible. In our analysis, LOCATE, as compared to BIANCA's global thresholding (with a global threshold of 0.99), did perform significantly worse at processing images with a low WMH load (see **Table 8**). Although having more true positives, which led to a very high sensitivity, LOCATE showed a 3-times higher FPR than BIANCA's global thresholding. Hence, all other metrics (DSC, OER, DER, H95 and FPR) showed worse outcomes for LOCATE compared to BIANCA's global thresholding. In addition, the WMH volumes received from LOCATE deviated significantly from the WMH volumes of the gold standards which is due to the massive number of false positives in LOCATE. With the global threshold in BIANCA this was not the case. Ling and colleagues (2018) validated BIANCA with different input modalities (FLAIR or FLAIR + T1w), with a cohort of patients with CADASIL using a semi-manually generated gold standard of 10 images per sequence. In their dataset, which contained an extremely high WMH load, they received a median DSC of 0.79 (our median 2D FLAIR = 0.560) for the 2D FLAIR + T1w images and a median DSC of 0.78 (our median DSC 3D FLAIR = 0.615) for the 3D FLAIR + T1w images. It was to be expected that the DSC in a dataset with a very high WMH load is higher than in a dataset with a low WMH load, as already described by several authors (Admiraal-Behloul et al., 2005; Anbeek et al., 2004; Khayati, Vafadust, Towhidkhah, & Nabavi, 2008; Sajja et al., 2006; P. Schmidt et al., 2012; Wack et al., 2012).

Importantly, the volumetric agreement for their 3D FLAIR + T1w, measured by the ICC, was very similar to ours with a single global threshold. However, we obtained higher values for the 2D FLAIR + T1w images. 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 subsets 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 6,** panel A and B). With a similar approach using the mean absolute WMH volume differences to the gold standard (FreeSurfer = 4.35, UBO 2D FLAIR = 2.00, BIANCA 2D FLAIR = 3.20), 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 ("insert **Supplementary Table 8** here").

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 (Griffanti et al., 2016). To our knowledge, BIANCA and LOCATE have not been validated with a longitudinal dataset so far.

Figure 6

Bland-Altman (Bland & Altman, 1986) *plots for WMH volume for the different algorithms (total WMH volume gold standard minus total WMH volume algorithm in mm^3). The x-axes contain mean WMH volumes, the y-axes contain absolute differences.*



4.2 Comparing the algorithms

The quality assessment of the algorithms is critically based on the gold standards. In order to prove construct validity, the different gold standards (T1w, 2D FLAIR, 3D FLAIR) were correlated amongst each other and with the respective outcomes of the algorithms (**"insert Supplementary Figure 2 here"**). The WMH volumes of the three gold standards correlated very strongly (all combinations: $r = 0.97$, $p < 0.05$) indicating a very high validity for our gold standards. However, evaluating how strong the different algorithm outputs correlated with the

gold standard WMH volumes, we found differences between the algorithms. UBO 3D FLAIR showed the highest correlations with the gold standard WMH volumes, followed by UBO 2D FLAIR, FreeSurfer, BIANCA 2D FLAIR, and – with the lowest correlations – BIANCA 3D FLAIR. 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. UBO Detector, on the other hand, has a built-in training dataset comprising ten 2D FLAIR images, which is used for analyses of both input modalities. Remarkably, the 2D and 3D FLAIR based WMH volumes of UBO Detector correlated strongly with the respective gold standard. Across modalities, the correlation of the algorithm outputs was very high between UBO 2D and 3D FLAIR ($r = 0.99$), while the correlation between BIANCA 2D and 3D FLAIR was clearly smaller ($r = 0.68$). Interestingly, the WMH volumes of FreeSurfer correlated also very high with the two UBO Detector outputs (2D FLAIR: $r = 0.92$, 3D: $r = 0.949$), but less strong with the two BIANCA outputs (2D FLAIR: $r = 0.75$, 3D: $r = 0.86$). In summary, these correlations indicate that the WMH volume segmentations by UBO Detector and FreeSurfer are better aligned as opposed to the BIANCA volume segmentation, which may be suggestive of segmentation errors.

Having a closer look on the validation metrics, we found that UBO Detector and BIANCA performed better than FreeSurfer in terms of volumetric agreement with the gold standard (ICC, OER). The WMH volumes extracted from FreeSurfer were generally smaller than the outputs of UBO Detector and BIANCA, and underestimated all gold standards (2D FLAIR, 3D FLAIR, T1w). We would like to emphasize that FreeSurfer is the only algorithm that has even underestimated its own gold standard. This clear volume underestimation is likely due to the fact that, sometimes, WMH appear isointense on T1w images, which also explains FreeSurfer's low false positive ratio and the high number of true negatives. The underestimation also affected other metrics (DSC, Sensitivity, OER and ICC), which were significantly worse for FreeSurfer compared to the other algorithms. Regarding overlap and resemblance agreement (DSC, DER, H95), BIANCA 3D FLAIR performed best. It scored significantly better than UBO Detector

regarding the H95, meaning that BIANCA matched the shape of the gold standards better. This can be an important advantage for research questions that regard lesion shape. A recent study, for example, has identified WMH shape as a marker to distinguish between patients with type-2 diabetes mellitus and a control group (de Bresser et al., 2018). Furthermore, BIANCA 3D but also BIANCA 2D FLAIR performed very accurate in terms of segmented percentage WMH voxels (confusion matrix) and average WMH volume compared to the other algorithms. On the flip side, both – BIANCA 3D and 2D FLAIR – showed the lowest correlations of WMH volume (a) with the Fazekas scores and (b) between the time points. In line with the latter, BIANCA 2D and 3D FLAIR showed the most «conspicuous intervals between two measurement points» and «subjects with conspicuous longitudinal data» compared to the other algorithms, and also as the only algorithm more «conspicuous intervals between two measurement points» than «subjects with conspicuous longitudinal data». This difference between intervals and subjects indicate that e.g. in BIANCA 2D FLAIR in 52 subjects at least two intervals between two measurement points were conspicuous, which in turn reflect the discovered visual zigzag pattern. The strongest WMH increases over time in literature were observed in subjects with a high WMH at baseline (Duering et al., 2013; Gouw et al., 2008; R. Schmidt et al., 2003). Little progression in punctate WMH but rapid progression in confluent WMH are reported (R. Schmidt, Seiler, & Loitfelder, 2016). However WMH can also cavitate to take on the appearance of lacunes and so they can also disappear (Shi & Wardlaw, 2016) for example after therapeutic intervention (Ramirez et al., 2016). Some studies report annual percentage increases in WMH volumes in the range between 12.5% and 14.4% in subjects with early confluent lesions, and 17.3% and 25.0% in subjects with confluent abnormalities (Duering et al., 2013; Sachdev, Wen, Chen, & Brodaty, 2007; R. Schmidt et al., 2003; van Dijk et al., 2008). Ramirez et al. (2016) summarized the progression rates of WMH volume in serial MRI studies in their Table 2, showing a wide variability of ranges. BIANCA 3D FLAIR showed a WMH volume increase range of 76.8% (mean) to 251.5% (mean + 1SD), and in 2D FLAIR images a range of 64.8% to 141.7% for 2D FLAIR images in one year. Comparing this progression of WMH volume growth with those described in the

1052 literature, and also considering that the WMH in this data set should increase only slightly, then
1053 these results again may point to segmentation errors, that influenced segmentation reliability.
1054 Having a closer look on the segmentation variability of the algorithms by means of the Bland-
1055 Altman plots **(see Figure 6)**, we observed that the limits of agreement are wider in BIANCA than
1056 in UBO Detector. Moreover, the 2D and 3D FLAIR plots show strong outliers (under- and
1057 overestimations). Interestingly, however, the single deviations in BIANCA seem to cancel each
1058 other out and result in a mean WMH volume that is very similar to the gold standard (see **"insert**
1059 **Supplementary Table 8 here**). Regardless of the WMH load, UBO Detector systematically
1060 underestimated the WMH volume in the 3D FLAIR images compared to the gold standard but
1061 BIANCA also showed a slight tendency to do so (see **Figure 6,** panel B). This effect, however,
1062 was not significant (**see Table 7**). From analyzing the validation metrics, we can conclude that
1063 with our dataset UBO Detector and FreeSurfer, as compared to BIANCA, performed more robust
1064 and consistent across time. Future research needs to evaluate if the segmentation errors, BIANCA
1065 produced with our dataset, occur also in the context of other datasets.
1066 Besides performance differences between algorithms, we also looked at an influence of input
1067 modality. Considering the findings, we obtained with our subset *«n162»,* for both algorithms –
1068 UBO Detector and BIANCA – that the segmented WMH volumes were significantly smaller
1069 when using 3D FLAIR compared to 2D FLAIR images. No such modality difference was
1070 apparent when comparing the 16 3D FLAIR and 2D FLAIR images between the gold standards
1071 within UBO Detector and within BIANCA (see **"insert Supplementary Table 6 here"**).
1072 It may be that this is specific to the subsample selected for the gold standard, and that differences
1073 would become apparent if we had manually segmented the 486 images from the *«n162»* subset.
1074 Findings with multiple sclerosis (MS) patients found that the number of lesions – manually
1075 segmented by radiologists – detected in 3D FLAIR compared to 2D FLAIR images were higher
1076 and therefore more WMH volume was detected (Paniagua Bravo et al., 2014; Polak, Magnano,
1077 Zivadinov, & Poloni, 2012; Tan et al., 2002). This aspect seems to be particularly obvious for
1078 brains with small WMH lesions, as they are more visible on 3D FLAIR images and especially in

1079 images with small slice thickness. Due to the statistically small sample, differences between
1080 modalities may be due to MRI acquisition. Furthermore, the KNN algorithm could also have an
1081 influence on the WMH volume output – or the interaction of acquisition and KNN algorithm.
1082 Hence, the potential effect of input modality in non-clinical samples needs to be further assessed
1083 in future studies.
1084 One general problem in the context of automated WMH lesion segmentation using FLAIR
1085 images is the incorrect inclusion of the septum pellucidum, the area separating the two lateral
1086 ventricles, in the output mask. This area appears hyperintense on FLAIR sequences, and
1087 therefore, looks very similar to WMH. When erroneously detected as WMH, the septum
1088 pellucidum enters the output volume as false positive region, which leads to an overestimation of
1089 the WMH volume. The UBO Detector developers (Jiang et al., 2018) also segmented the septum
1090 pellucidum in their gold standard (see their supplementary Figure 1b). Since they already fed
1091 their algorithm with this false positive information, it was to be expected that UBO Detector
1092 would also segment the septum pellucidum in our data, which may have caused the worse DER
1093 compared to FreeSurfer and BIANCA. Interestingly, however, also the BIANCA and LOCATE
1094 outputs included some false positives in the septum pellucidum although the algorithm was
1095 trained with a customized training dataset in which this area was not segmented.
1096

1097 4.3 Strength and limitations

1098 The main strengths of this study are the validation and comparison of three freely available
1099 algorithms using a large longitudinal dataset of cognitively healthy subjects with a low WMH
1100 load. Furthermore, we used fully manually segmented WMH as references for all three MRI
1101 sequences. For the 3D and 2D FLAIR modalities, all three operators segmented the same images
1102 in order to determine the inter-operator agreement. Scans with smaller WMH load lead to a lower
1103 DSC than scans with a high WMH load (Wack et al., 2012). This is because images with high
1104 lesion loads are «easier» for operators to achieve high DSC values than images with low lesion

loads since images with high lesion loads usually show large lesions that the operators can easily agree on. Moreover, the volume to lesion ratio in brains with high WMH load is smaller compared to brains with low WMH load. We achieved good to excellent inter-operator agreements (Caligiuri et al., 2015; Dadar et al., 2017) among the three operators, regardless of the low WMH load. Our results on the OER and DER further indicate that the errors were mainly due to edges rather than to missing voxels. So far, there is no general convention for interpreting these measures, but our results are well in line with the study of Wack et al. (2012), who reported similar values based on two operators (OER = 0.41, DER = 0.15). Supporting the high quality of our manual segmentations, we found an excellent (Cicchetti, 1994) reliability for the total volumetric agreement between the measurements of the 3D and 2D FLAIR images (ICC: 3D FLAIR mean = 0.96; 2D FLAIR = 0.82). There were no significant mean WMH volume differences between the three gold standards in all sequences ("insert **Supplementary Table 6** here"). Besides the manual segmentations, the whole dataset (*N* = 800) was rated by the three operators using the Fazekas scale, so that inter-operator reliability overall and between operators could be calculated. The rating comparisons of all three operators resulted in substantial to almost perfect agreement (Landis & Koch, 1977). Thus, the Fazekas scores can be considered as a reliable gold standard to cross-validate WMH volumes extracted with the automated algorithms. Finally, a longitudinal subset with 162 data points containing all three modalities was created. The limitation of this study is that it is only one sample with homogeneous participants, which on the other hand makes the data comparable. Future studies will determine how well these results generalize to other studies, scanners, sequences, heterogeneous datasets with clinical participants.

4.4 Usability of the algorithms

Given that the algorithms for WMH extraction are usually not implemented by trained programmers, usability is an important issue to also mention here.

FreeSurfer has not been specifically programmed for WMH detection, but is a tool for extensive analysis of brain imaging data. Because of all the other parameters FreeSurfer outputs besides WMH volume, the processing time is very long (many hours per session). The FreeSurfer output comprises the total WMH volume and the total non-WMH volumes (grey matter).

UBO Detector has been specifically programmed for WMH detection. The UBO Detector algorithm has been trained with a «built-in» training dataset. Theoretically, it is possible to train the algorithm using a previously manually segmented gold standard. However, this procedure does only work within the Graphic User Interface (GUI) in DARTEL space, and is very time-consuming. The output from UBO Detector is well structured and contains among others the WMH volume and the number of clusters for total WMH, PVWMH, DWMH as well as WMH volumes per cerebral lobe. For subset 2 the WMH extraction for the whole process incl. pre- and postprocessing, took approx. 14 min per brain with the computing environment specified in the methods section. For subset 3 the WMH extraction took 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 validation and comparison of three freely available methods for automated WMH segmentation using a large longitudinal dataset of cognitively healthy subjects with a 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 it provides good agreement with the Fazekas scores and delivers robust and constant data in terms of WMH volume over time, its use as a control variable is conceivable. BIANCA received in general very good cross-sectional accuracy metrics but the longitudinal data suggest that it generates false positive WMH volumes in certain areas of the brain. Hence, about 30% of the automatically generated segmentations are not usable without great manual effort. The random false positives, which frequently occurred in frontal brain areas, could thus also lead to random results in e.g. correlations of WMH and cognitive functions. 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 DSC, DER and H95, would be desirable to be considered as a true replacement for manual segmentation of WMH.

6 Acknowledgements

We acknowledge all participants. We are thankful to Susanne Wehrli for her help on segmenting WMH masks used in our study.

7 Conflicts of Interest

The authors declare no conflicts of interest.

8 Information Sharing Statement

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)

References

Admiraal-Behloul, F., van den Heuvel, D. M. J., Olofsen, H., van Osch, M. J. P., van der Grond, J., van Buchem, M. A., & Reiber, J. H. C. (2005). Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *Neuroimage*, *28*(3), 607–617. doi:10.1016/j.neuroimage.2005.06.061

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

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

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

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

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 Bresser, J., Kuijf, H. J., Zaanen, K., Viergever, M. A., Hendrikse, J., Biessels, G. J., & Utrecht Vascular Cognitive Impairment Study Group. (2018). White matter hyperintensity shape and location feature

analysis on brain MRI; proof of principle study in patients with diabetes. *Scientific Reports*, *8*(1), 1893. doi:10.1038/s41598-018-20084-y

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

Duering, M., Csanadi, E., Gesierich, B., Jouvent, E., Hervé, D., Seiler, S., … Dichgans, M. (2013). Incident lacunes preferentially localize to the edge of white matter hyperintensities: insights into the pathophysiology of cerebral small vessel disease. *Brain: A Journal of Neurology*, *136*(Pt 9), 2717–2726. doi:10.1093/brain/awt184

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

Gouw, A. A., van der Flier, W. M., Fazekas, F., van Straaten, E. C. W., Pantoni, L., Poggesi, A., … LADIS Study Group. (2008). Progression of white matter hyperintensities and incidence of new lacunes over a 3-year period: the Leukoaraiosis and Disability study. *Stroke*, *39*(5), 1414–1420. doi:10.1161/STROKEAHA.107.498535

Griffanti, L., Jenkinson, M., Suri, S., Zsoldos, E., Mahmood, A., Filippini, N., … Zamboni, G. (2018). Classification and characterization of periventricular and deep white matter hyperintensities on MRI: A study in older adults. *Neuroimage*, *170*, 174–181. doi:10.1016/j.neuroimage.2017.03.024

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

Khayati, R., Vafadust, M., Towhidkhah, F., & Nabavi, M. (2008). Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model. *Computers in Biology and Medicine*, *38*(3), 379–390. doi:10.1016/j.compbiomed.2007.12.005

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

Paniagua Bravo, Á., Sánchez Hernández, J. J., Ibáñez Sanz, L., Alba de Cáceres, I., Crespo San José, J. L., & García-Castaño Gandariaga, B. (2014). A comparative MRI study for white matter hyperintensities

detection: 2D-FLAIR, FSE PD 2D, 3D-FLAIR and FLAIR MIP. *The British Journal of Radiology*, *87*(1035), 20130360. doi:10.1259/bjr.20130360

Polak, P., Magnano, C., Zivadinov, R., & Poloni, G. (2012). 3D FLAIRED: 3D fluid attenuated inversion recovery for enhanced detection of lesions in multiple sclerosis. *Magnetic Resonance in Medicine*, *68*(3), 874–881. doi:10.1002/mrm.23289

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

Sachdev, P., Wen, W., Chen, X., & Brodaty, H. (2007). Progression of white matter hyperintensities in elderly individuals over 3 years. *Neurology*, *68*(3), 214–222. doi:10.1212/01.wnl.0000251302.55202.73

Sajja, B. R., Datta, S., He, R., Mehta, M., Gupta, R. K., Wolinsky, J. S., & Narayana, P. A. (2006). Unified approach for multiple sclerosis lesion segmentation on brain MRI. *Annals of Biomedical Engineering*, *34*(1), 142–151. doi:10.1007/s10439-005-9009-0

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, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., … Mühlau, M. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *Neuroimage*, *59*(4), 3774–3783. doi:10.1016/j.neuroimage.2011.11.032

Schmidt, R., Enzinger, C., Ropele, S., Schmidt, H., Fazekas, F., & Austrian Stroke Prevention Study. (2003). Progression of cerebral white matter lesions: 6-year results of the Austrian Stroke Prevention Study. *The Lancet*, *361*(9374), 2046–2048. doi:10.1016/s0140-6736(03)13616-1

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

Tan, I. L., Pouwels, P. J. W., van Schijndel, R. A., Adèr, H. J., Manoliu, R. A., & Barkhof, F. (2002). Isotropic 3D fast FLAIR imaging of the brain in multiple sclerosis patients: initial experience. *European Radiology*, *12*(3), 559–567. doi:10.1007/s00330-001-1170-8

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

van Dijk, E. J., Prins, N. D., Vrooman, H. A., Hofman, A., Koudstaal, P. J., & Breteler, M. M. B. (2008). Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences: Rotterdam Scan study. *Stroke*, *39*(10), 2712–2719. doi:10.1161/STROKEAHA.107.513176

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