Medical Image Analysis 38 (2017) 50–64

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Longitudinal segmentation of age-related white matter hyperintensities

Carole H. Sudre[a,b,*], M. Jorge Cardoso[a,b], Sebastien Ourselin[a,b], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] *Translational Imaging Group, Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK*
[b] *Dementia Research Centre, UCL Institute of Neurology, University College London, London, UK*

ARTICLE INFO

Article history:
Received 1 August 2016
Revised 13 February 2017
Accepted 15 February 2017
Available online 24 February 2017

Keywords:
White matter hyperintensities
Segmentation
Longitudinal

ABSTRACT

Although white matter hyperintensities evolve in the course of ageing, few solutions exist to consider the lesion segmentation problem longitudinally. Based on an existing automatic lesion segmentation algorithm, a longitudinal extension is proposed. For evaluation purposes, a longitudinal lesion simulator is created allowing for the comparison between the longitudinal and the cross-sectional version in various situations of lesion load progression. Finally, applied to clinical data, the proposed framework demonstrates an increased robustness compared to available cross-sectional methods and findings are aligned with previously reported clinical patterns.

1. Introduction

White matter hyperintensities (WMH), also known as leukoaraiosis, as observed in FLuid Attenuated Inversion Recovery (FLAIR), T2-weighted (T2) and proton density weighted (PD) magnetic resonance (MR) images are widely observed in the ageing population. The abnormal signal, explained by a change in the fat/water ratio, reflects a damage to the white matter. Hypotheses related to deleterious changes in the blood supply and in the blood brain barrier (Wardlaw et al., 2013) have been put forward to explain the occurrence of such damage, and cardiovascular risk factors such as hypertension have been shown to be associated to the WMH burden (Abraham et al., 2015; Vuorinen et al., 2011). Furthermore, such lesions have been linked with cognitive impairment, in particular with respect to processing speed and executive function (Prins and Scheltens, 2015; Wakefield et al., 2010).

To further assess potential causality effects between lesion burden and clinical outcome, new emphasis has been given to longitudinal studies of lesion load and cognitive assessment (Schmidt et al., 2005). In normal ageing, increase in the lesion volume with time was observed with a higher rate of change correlated with more severe baseline lesion volume (Pantoni and The LADIS Study group, 2011). For a normal population, progression in leukoaraiosis has been related to motor decline (Silbert et al., 2008), and cognitive disabilities (Schmidt et al., 2005) as well as memory impairment (Gunning-Dixon and Raz, 2000). Additionally, lesion burden at baseline has been associated with faster cognitive decline in Alzheimer's disease (AD), mild cognitive impairment (MCI) and normal populations (Carmichael et al., 2010).

The evaluation of WMH progression, however, remains difficult. In many cases, visual rating scales are used to assess the increase in severity of the lesion burden (Gouw et al., 2008). Most of them have however been developed for cross-sectional studies and are difficult to utilise in longitudinal cases due to the lack of sensitivity to change (Schmidt et al., 2005). Specific progressive rating scales have been proposed to alleviate this drawback (Prins et al., 2004), but volumetric measurements appear to allow for more accurate group differentiation (Pantoni and The LADIS Study group, 2011). Even when using semiautomatic segmentation methods for volume assessment (Schmidt et al., 2005) instead of performing the segmentation manually, the process remains time-consuming and the strategy of looking at images back-to-back can introduce bias

[*] Corresponding author at: Translational Imaging Group, Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, 8th Floor Malet Place Engineering Building, 2 Malet Place, WC1E 7JE, London UK

E-mail address: carole.sudre.12@ucl.ac.uk (C.H. Sudre).

[1] Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.media.2017.02.007

(Schmidt et al., 2005). Therefore, longitudinal, robust automatic lesion segmentation solutions are greatly needed.

Even though imaging time points can be considered independently when automatically measuring the volume of WMH in longitudinal studies (Carmichael et al., 2010), it has been shown that considering the time points separately within subject introduced an additional source of variability in the results (Elliott et al., 2013). Accounting for the structural similarities between time points, or relating the information from one time point to others may increase the robustness of the method.

The problem of longitudinal lesion assessment is of great interest in other fields of neuroimaging such as multiple sclerosis (MS), and various methods have been designed to assess longitudinal lesion change. This issue is especially sensitive in MS, in which the lesion load progression is non-monotonic. Methods relying on the analysis of the differences between registered serial images, as in Rey et al. (2002), may be hindered by other volumetric changes occurring between the time points. In studies with long-term follow-up, in which the drop-off rate can be high (*e.g.* in age-related studies), being able to handle different numbers of time points is an additional challenge.

In the context of age-related WMH, the progressive nature of the damage can be taken as an argument to consider consecutive image pairs as in Bosc et al. (2003). However, noise and artefacts, prevalent in aging or in the demented population, may affect methods based on direct comparison; other solutions based on image averaging and model building may be advantageous. For instance, the use of average images to guide the processing of longitudinal data has been promoted in Reuter et al. (2012).

The solution developed in this work first consists in creating a longitudinal intra-subject average (Section 2.1), followed by the estimation of an appropriate joint Gaussian mixture model (GMM) (Section 2.2) that will finally be used to constrain the lesion segmentation at each time point (Section 2.3). The main assumption of this work is that all time points can be diffeomorphically mapped to a subject-specific mean appearance.

To assess the relevance of the proposed technique for the study of WMH progression, a longitudinal lesion simulator was developed (Section 3.1) so as to test the method with various longitudinal patterns and lesion loads. A surrogate clinical validation was performed using data from the Alzheimer's disease Neuroimaging Initiative (ADNI) to test whether documented cross-sectional as well as longitudinal findings reported in the literature could be reproduced.

2. Method

In the following the subscript τ denotes a specific time point and GW the groupwise average appearance model. Prior to the construction of the average, an expectation maximisation (EM) algorithm with outlier detection and bias field correction is performed on each individual time point. The intensities $\mathbf{Y}_\tau$ are the resulting log-transformed, normalised and bias field corrected intensities of the skull-stripped images. With N the number of voxels and D the number of modalities, image intensities are vectorised into $Y^{(d)} = \{y_{d1}, \cdots, y_{dn}, \cdots y_{dN}\}$ with y_{dn} the intensity at voxel n of modality d, so that

$$\mathbf{Y} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(D)} \end{pmatrix}.$$

2.1. Longitudinal intra-subject average

In order to build the average appearance model, two main components linking the individual images to the average space are needed: a spatial transformation and an intensity transformation. An intensity matching between images is needed to account for changes in contrast, MR scanning variations and some artefacts. These transformations are obtained through an iterative process, proved to limit bias towards a specific time point. In order to avoid unrealistic spatial deformations, affine transformations roughly aligning the images are first applied before considering non-rigid transformations to obtain the final spatial transformations $T_{\tau\,\rightarrow\,\mathrm{GW}}$. At each iteration, the intensities of the images spatially transformed to the GW space are mapped to the intensities of the current average image using a polynomial fit of degree 2 for each modality used. More formally, the intensity mapping and the resulting mapping coefficients $h_\tau^{(d)}$ for one modality d can be expressed as

$$\underset{h_\tau^{(d)}}{\operatorname{argmin}} \parallel A\Big(T_{\tau\rightarrow\mathrm{GW}}(Y_\tau^{(d)})\Big)\cdot h_\tau^{(d)} - Y_{\mathrm{GW}}^{(d)} \parallel^2$$

where $A(T_{\tau\rightarrow\mathrm{GW}}(Y_\tau^{(d)}))$ is the polynomial matrix transformation of $T_{\tau\rightarrow\mathrm{GW}}(Y_\tau^{(d)})$ such that

$$A(Y) = \begin{pmatrix} 1 & y_1 & y_1^2 \\ \vdots & \vdots & \vdots \\ 1 & y_N & y_N^2 \end{pmatrix}.$$

The steps to create an average appearance model are:

Step 1 Register each of the individual time points to the current average image.
Step 2 Map the intensities of each resampled image to the current average image using a polynomial fit of degree 2.
Step 3 Average all resampled and intensity transformed images to create the new current average image.
Step 4 Go back to step 1.

With this set up the loop is performed five times: the first iteration consists in the estimation of a rigid transformation followed by two affine transformations before allowing for a non-rigid registration at the last two iterations.

2.2. Model selection

After creating the average appearance model, patient-specific tissue priors and brain mask are obtained using the GIF (Geodesic Information Flow) pipeline developed in Cardoso et al. (2015). In this method, label-fusion is used to generate subject specific tissue priors (**A**) by propagating pre-segmented templates and fusing them locally according to Cardoso et al. (2015). Using the priors and brain mask as inputs, BaMoS (Sudre et al., 2015) is used to model the data according to a three-level Gaussian mixture. The first level segments inliers from outliers observations while the anatomical tissue information is introduced at the second level so that each of the inlier and outlier tissue classes is modelled by a Gaussian mixture at the third level of the model. The final distribution model is then expressed as

$$f(\mathbf{Y}|\boldsymbol{\Xi}_{\mathbf{K}}) = \prod_{n=1}^{N}\sum_{l\in I,O}\sum_{j=1}^{J}\sum_{k=1}^{K_{l_j}+1}\pi_{nl_{j_k}}\mathcal{M}\Big(\mathbf{y}_n|\theta_{l_{j_k}}\Big)$$

where $\pi_{nl_{j_k}}$ are the spatially varying weights in the mixture obtained by multiplying the class mixing proportions, and the inlier and tissue at the previous levels, l refers to the segmentation between inliers (I) and outliers (O), j to the anatomical classes and k to the individual Gaussian components. The notation $\mathbf{K}$ is used to encompass the model complexity (number of components for each tissue class K_{l_j}), while $\boldsymbol{\Theta}$ gathers the model parameters of each individual component (mixture weight $w_{l_{j_k}}$, mean $\boldsymbol{\mu}_{l_{j_k}}$ and covariance matrix $\Lambda_{l_{j_k}}$). Finally $\boldsymbol{\Xi}$ is defined as $\{\mathbf{K}, \boldsymbol{\Theta}, \mathbf{A}, \mathbf{B}\}$, $\mathbf{B}$ being the inlier/outlier atlases. In order to improve the sensitivity of BaMoS to outlier lesions, an additional step is included in the initialisation phase of the algorithm. Instead of the flat outlier priors used in the original BaMoS paper, those are replaced after the first EM convergence with a spatially varying outlierness map, also known as a typicality map (Van Leemput et al., 2001). This typicality map is estimated using the formulation presented in Van Leemput et al. (2001) with κ=3. The typicality value for a given voxel n follows the expression:

$$t_n = \sum_{j=1}^{J} p_{nj} \frac{\mathcal{G}\left(\mathbf{y}_n|\theta_{I_j}\right)}{\mathcal{G}\left(\mathbf{y}_n|\theta_{I_j}\right) + \dfrac{1}{\sqrt{(2\pi)^D|\Lambda_{I_j}|}\exp\left(-\frac{1}{2}\kappa^2\right)}}$$

where $p_{nj} = p_{nI_j} + p_{nO_j}$

2.3. Constraint over individual time points

Once a model for the longitudinal intra-subject average image has been obtained, it can be used to constrain the lesion segmentation of each time point. First, the anatomical subject-specific statistical atlases are transformed to the space of each time point using the backward transformations $T_{\mathrm{GW}\rightarrow\tau}$ obtained during the averaging process. The groupwise model parameters $\boldsymbol{\Theta}^{\mathrm{GW}}$ are then used as priors over the model parameters for each time point whose intensities are mapped to the average appearance model. The model structure $\mathbf{K}^{\mathrm{GW}}$ estimated for the average image is also preserved. Using $\boldsymbol{\Theta}^{\mathrm{GW}}$ as priors and $\mathbf{K}^{\mathrm{GW}}$ as the model, the individual time point model parameters are estimated through an EM algorithm. Priors over the means are introduced as normal distributions while Inverse-Wishart distributions are chosen as priors for the covariance. As such, the expectation step is the same as in BaMoS but the M-step consists of maximizing at iteration t the following expectation $\mathbb{E}$, considering $\mathbf{Z}_\tau$ the hidden data labels

$$\mathbb{E}_{f\left(\mathbf{Z}_\tau|\hat{\mathbf{Y}}_\tau, \boldsymbol{\Xi}^{\tau\,(t-1)}_{K^{\mathrm{GW}}}, \hat{\mathbf{h}}_\tau\right)} f\left(\hat{\mathbf{Y}}_\tau, \mathbf{Z}_\tau|\boldsymbol{\Xi}^{\tau}_{K^{\mathrm{GW}}}, \hat{\mathbf{h}}_\tau\right) \cdot f\left(\boldsymbol{\Xi}^{\tau}_{K^{\mathrm{GW}}}|\boldsymbol{\Xi}^{\mathrm{GW}}_{K^{\mathrm{GW}}}\right).$$

The distribution $f\left(\boldsymbol{\Xi}_{K^{\mathrm{GW}}}|\boldsymbol{\Xi}^{\mathrm{GW}}_{K^{\mathrm{GW}}}\right)$ in which the script τ has been dropped for notation convenience is expressed as

$$f\left(\boldsymbol{\Xi}_{K^{\mathrm{GW}}}|\boldsymbol{\Xi}^{\mathrm{GW}}_{K^{\mathrm{GW}}}\right) = \prod_{l\in I,O}\prod_{j=1}^{J}\prod_{k=1}^{K_{l_j}}\left[\mathcal{G}\left(\boldsymbol{\mu}_{l_{j_k}}|\boldsymbol{\mu}^{\mathrm{GW}}_{l_{j_k}}, \Lambda^{\mathrm{GW}}_{l_{j_k}}\right)\mathcal{IW}\left(\Lambda_{l_{j_k}}|\tilde{N}\Lambda^{\mathrm{GW}}_{l_{j_k}}, N\right)\right],$$

where $\mathcal{G}$ refers to a normal distribution and $\mathcal{IW}$ to an Inverse-Wishart distribution with $\tilde{N} = N + D + 1$. With $\Omega^{(t)}_{l_{j_k}}$ being the weighted covariance matrix, incorporating these distributions into the maximisation process leads to the following update equations:

$$\boldsymbol{\mu}^{(t)}_{l_{j_k}} = \left(\sum_{n=1}^{N} p^{(t)}_{nl_{j_k}} M_{\hat{\mathbf{h}}}(\hat{\mathbf{y}}_n)\Lambda^{-1(t-1)}_{l_{j_k}} + \boldsymbol{\mu}^{\mathrm{GW}}_{l_{j_k}}\Lambda^{\mathrm{GW}^{-1}}_{l_{j_k}}\right) \cdot \left(\sum_{n=1}^{N} p^{(t)}_{nl_{j_k}}\Lambda^{(t-1)^{-1}}_{l_{j_k}} + \Lambda^{\mathrm{GW}^{-1}}_{l_{j_k}}\right)^{-1}$$

$$\Lambda^{(t)}_{l_{j_k}} = \frac{\tilde{N}\Lambda^{\mathrm{GW}}_{l_{j_k}} + \sum_{n=1}^{N} p^{(t)}_{nl_{j_k}}\Omega^{(t)}_{l_{j_k}}}{\sum_{n=1}^{N} p^{(t)}_{nl_{j_k}} + \tilde{N}}.$$

Following this optimisation for each time point, the lesion segmentation can be obtained. It consists here of the selection of a subset of the outliers weighted with respect to the characteristics of the WM inlier distribution : first, outlier voxels with an inlier or hypo-intense FLAIR Mahalanobis distance with respect to the healthy white matter are excluded; lesion clusters that fall outside of the white matter mask are also excluded. This mask is obtained by excluding regions obtained from the label propagation framework that cannot plausibly correspond to WM lesions such as the ventricles and the cortical ribbon. To illustrate the longitudinal framework, Fig. 1 presents a graphical representation of the longitudinal segmentation process.

3. Validation on simulated data

In order to assess the validity and sensitivity of the longitudinal framework developed in this work, both synthetic and clinical data were used. Synthetic data allows for a ground truth comparison and was designed to assess the sensitivity to change of the tested algorithms.

3.1. Lesion simulator

3.1.1. Image production

In line with the synthetic image building detailed in Jack Jr et al. (2001), two sets of data are used to simulate lesions: a receiving set, comprised of subjects with minimal to no WMH, and a donating set, comprised of images with non zero WMH lesion load and their associated probabilistic lesion segmentation. The process of simulating lesions involves spatially and intensity transforming the lesions from the donating set to the receiving set. For the receiving set, T1-weighted and FLAIR MRI scans of the ADNI database were used, with the FLAIR image affinely registered to the T1 space resulting in 1 mm^3 isotropic images. To ensure that lesions are contained within the WM, lesion maps are first propagated under the constraint that they fall within the WM. These lesions are then shrunk to simulate different longitudinally consistent lesion loads. To simulate shrunken lesions, the initial propagated WMH load is modified by thresholding the probabilistic lesion segmentation L at a certain value X, followed by a normalisation step, i.e. $L_S = (L - X)/X \;\forall L > X$, with L being the original lesion probability per voxel. The value of X is chosen to produce an exact reduction in WMH volume of D. As L_{new} can contain hard edges, L_S is then smoothed with a Gaussian filter and, due to the non-volume-preserving nature of the Gaussian smoothing process, re-mapped to have an exact reduction in WMH volume of D through a piecewise linear transformation. Defining p_b such that

$$\sharp L_S|L_S > p_b = \sharp L|L > 0.5 - D,$$

the following system to define the two linear mapping is solved:

$$\left.\begin{array}{rl} b_1 & = 0 \\ a_1 \cdot p_b + b_1 & = 0.5 \end{array}\right\} \text{if } L_S < p_b$$

$$\left.\begin{array}{rl} a_2 \cdot p_b + b_2 & = 0.5 \\ a_2 + b_2 & = 1 \end{array}\right\} \text{if } L_S > p_b$$

Since p_{new} is only a probability and not an actual intensity, p_{new} is transformed into an intensity map by drawing samples from a Gaussian distribution with parameters given by the lesion distribution of the donating set.

To account for variation over time of scanner characteristics and subject positioning when simulating lesion evolution and atrophy random bias field and rigid transformations are applied to the images. The bias field, modelled as a linear combination of polynomial basis functions is obtained by randomly choosing the linear coefficients. Given a probabilistic lesion map L, the lesion intensities G sampled from a Gaussian distribution, the log-transformed bias field BF, the rigid transformation R and the initial image I, the

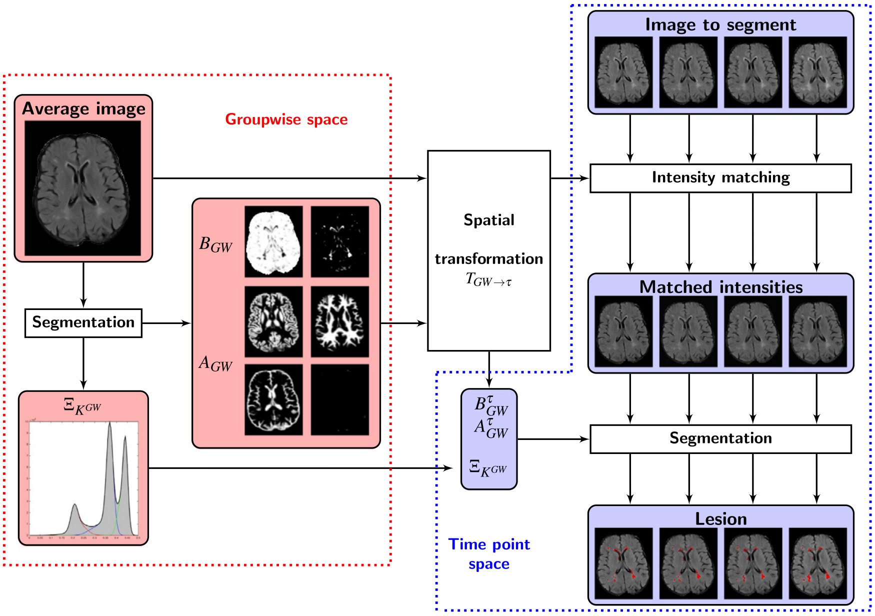

Fig. 1. Diagram of the longitudinal segmentation process.

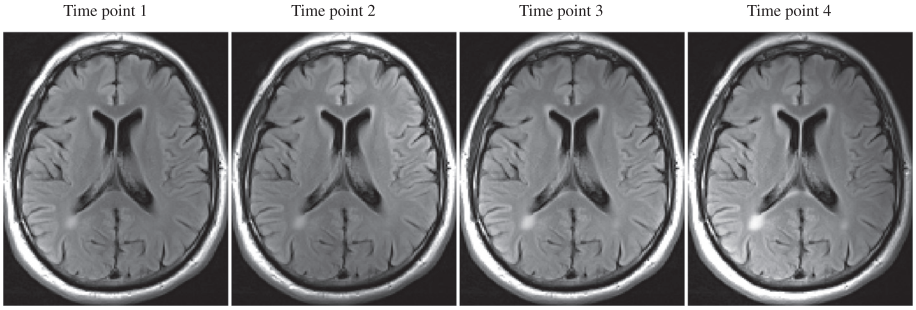

Fig. 2. Results of the lesion simulator after application of the random bias field for four time points with no rigid transformation applied. For realism purposes of increased lesion burden the time points are reversed compared to their order of simulation.

final simulated image S can be expressed at voxel n by

$$S_n = R(\exp(BF_n) \cdot (L_n G_n + (1 - L_n) I_n)).$$

In this work, a standard deviation of 1 mm was used for the Gaussian smoothing. An example of the outcome of the lesion simulator is presented in Fig. 2 on which the same slice of the FLAIR image (before rigid transformation) is presented at four time points of the lesion progression.

3.1.2. Simulated evolution patterns

Different patterns of WMH progression with differing lesion loads and time points were simulated. The database used for WMH simulation comprised of 5 donating images and 17 receiving images. 4 patterns were simulated using the following implementation:

Linear_500: Linear reduction of 500 mm^3 per step, spanning 6 time points.

Linear_750: Linear reduction of 750 mm^3 per step, spanning 4 time points.

NonLinear_5: Non-linear reduction of 5% per step, spanning 6 time points.

NonLinear_15: Non-linear reduction of 15% per step, spanning 4 time points.

Although the progression are implemented by load shrinkage, for clinical realism and illustration purposes, the time points are then reordered to simulate a progressive increase in lesion load.

For each of these progression schemes (denoted Slope if not modified), two additional plateauing patterns were added to test for longitudinal bias:

Flat_High 1 time point with highest load was added to form a high plateau.

Flat_Low 2 time points with lowest load were added to form a low plateau.

Then, to simulate treatment effect, composite patterns were created using the simulated linear patterns in order to simulate changes in the slope:

Treatment One increase step of 750 mm^3 followed by two steps with an increase of 500 mm^3.

No treatment 3 steps with an increase of 750 mm^3 per step.

Fig. 3a plots an example of the four typical simulated evolution paths with similar maximum loads, different minimum loads and their associated plateauing versions while Fig. 3b presents an example of the modelled combination of linear patterns to simulate the treatment effect case.

Finally, in order to assess situations where brain atrophy occurs concomitantly to lesion progression, atrophy was further simulated with a linear lesion progression of 750 mm^3 per step. To do so, the progressive deformations observed in an AD subject were applied sequentially to the receiving image. In order to maintain the realism of the lesion maps, the zone

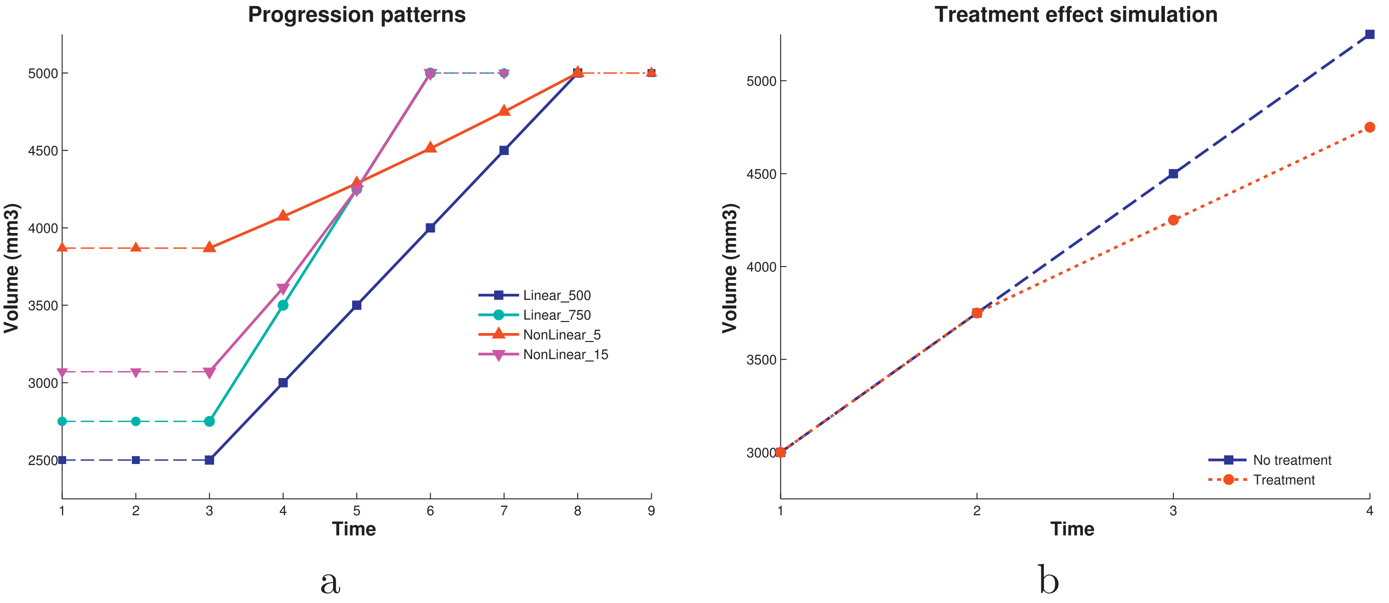

Fig. 3. Left) Example of the four simulated evolution patterns. The dashed horizontal lines represent the plateauing experiments at either high (Flat_High) or low (Flat_Low) load. Right) Example of the combination of two linear patterns to simulate a treatment related change.

Table 1
Summary of the ground truth volumes (Lesion probability map – intersection of baseline segmentations) across the different evolution patterns.

	NonLinear_5	Linear_750	Linear_500	NonLinear_15
Mean	2871	2793	2645	2510
SD	2519	2619	2594	2307
Median	2881	2542	2379	2206
IQR	[467 4040]	[314 4176]	[188 3941]	[401 3745]

of allowed lesion evolved accordingly. The NiftyReg package (https://sourceforge.net/p/niftyreg/git/ci/dev/tree/) was used for all registration and resampling operations. Since the changes across time points are the most severe in this specific series, a test of robustness with respect to the order in which the images are considered to form the average was performed considering the forward and backward series of time points.

3.2. Segmentation assessment

The longitudinal framework (Long) was compared to the cross-sectional application of BaMoS performed both in its original cross-sectional version (Cross) (Sudre et al., 2015) and its sensitivity enhanced variant (Cross+), i.e. when using the typicality map to estimate the outlier atlas. As the receiving images used in the simulation can contain trace amounts of lesions, the region where all methods agreed in the presence of lesions for the time point with minimal lesion load was excluded from the analysis, both in terms of volume and overlap. Ground truth (GT) lesion segmentations are thus the simulated lesion probability maps corrected for the baseline lesion segmentation intersection of all given methods.

Statistics of the ground truth volumes are presented in Table 1.

Those corrected differences were finally compared in terms of Dice score coefficient (DSC), true positive rate (TPR) and average distance (AvDist) as defined in Styner et al. (2008). Due to the simulation process, images cannot be considered as independent for the same subject. Therefore, statistics were calculated over the mean per subject in each pattern. The origin of the errors was further investigated differentiating false positive (FP) and false negatives (FN), outline (OE) and detection error (DE) as defined in Wack et al. (2012). Definition of the assessment measures can be found in Appendix A.

The publicly available toolbox LST included in SPM was used as a point of external comparison in the experiment relative to the treatment effect and in the sequence where atrophy was simulated. This method, proposed in Schmidt et al. (2012) develops a lesion growing model based on the thresholding of outlier beliefs maps. This threshold has to be chosen by the user, with a default of 0.30 and this choice is denoted LST-d. The value of 0.25 has been considered as adequate when dealing with ageing populations (Manjón et al., 2010) and is therefore the second value chosen for comparison and denoted LST-a. Recently a new pipeline called LPA that does not require any user interaction has been included in the toolbox. This configuration was also tested as additional point of comparison. Additionally, the Lesion-TOADS algorithm, available as a plugin to the medical image analysis software MIPAV was further used for comparison purposes. This method uses fuzzy C-means in a framework ensuring topological consistency and correcting for bias field (Shiee et al., 2010). Aligned skull-stripped T1 and FLAIR images were provided as input to the algorithm using the same mask and inter-sequence alignment as for Cross, Cross+ and Long to ensure comparability.

3.3. Results

3.3.1. Evolution patterns

The assessment across the evolution patterns are presented in Table 2. With respect to the differences in DSC better scores were observed for patterns with lower ranges of change and higher median load (NonLinear_5) For the lesion loads allowing a complete evolution for the four evolution patterns, the slopes of extracted volumes obtained for the three methods and the ground truth are given in Table 3. Slopes were obtained using a mixed effects model with random slope and intercept. Note that the rigid transformation may modify the actual volume of lesion and thus explains why the ground truth slopes are slightly different from expected in the case of linear transformations. As an approximation, non linear evolutions are also linearly modelled.

Table 2
Segmentation assessment table for the longitudinal framework according to the different strategies of evolution. Results are given in the form median [IQR], where the median are calculated across subjects on the scores average over time points.

	Linear_500	Linear_750	NonLinear_5	NonLinear_15
DSC	0.65	0.66	0.66	0.64
	[0.28 0.77]	[0.34 0.76]	[0.41 0.76]	[0.41 0.74]
TPR	0.83	0.80	0.81	0.79
	[0.66 0.91]	[0.57 0.88]	[0.65 0.88]	[0.67 0.85]
AvDist	2.07]	1.89	1.93	2.05
	[1.00 9.83]	[1.06 9.69]	[0.96 5.46]	[1.11 6.62]
OE/TotF	0.81	0.77	0.82	0.80
	[0.49 0.90]	[0.53 0.88]	[0.67 0.90]	[0.57 0.88]
OEFP/FP	0.72	0.71	0.74	0.69
	[0.41 0.86]	[0.47 0.85]	[0.55 0.87]	[0.50 0.85]
OEFN/FN	0.97	0.93	0.97	0.94
	[0.88 1.00]	[0.78 0.97]	[0.91 1.00]	[0.90 0.97]
FP/TotF	0.83	0.79	0.77	0.76
	[0.60 0.92]	[0.59 0.86]	[0.56 0.87]	[0.59 0.82]

Acronyms expansion: DSC - Dice Similarity Coefficient; TPR - True Positive Rate; AvDist - Average Distance; FP/TotF - Proportion of false positives in the total of error; OE/TotF - Proportion of outline error in the total error; OEFP/FP - Proportion of false positive outline error in the false positives; OEFN/FN - Proportion of false negative outline error in the false negatives.

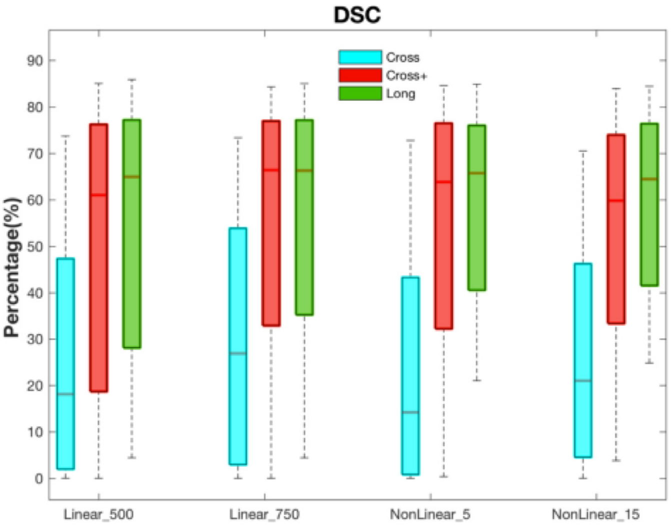

Fig. 4. Comparison of the DSC distribution between the three methods across the different evolution patterns. Borders of the boxes represent the 25th and 75th percentile while the thick line in each box corresponds to the median. Whiskers are limited to the 5th and 95th percentiles.

3.3.2. Plateauing bias evaluation

Possible bias introduced by a flat WMH load at the lowest (Flat_Low) or at the highest (Flat_High) end of the progression period was evaluated on the common time points between the sets. The results for this experiment are presented in Table 4 emphasising the stability of the method when including plateauing time points. Additionally the stability between time points on the plateauing regions was evaluated using Lin concordance and the percentage of change observed evaluated for the three presented methods. Results are presented in Table 5 and show a stronger concordance for Long compared to Cross and Cross+ in both plateauing situations (high and low loads). Furthermore, the percentage difference between the detected volumes at plateauing region was non-significantly different from 0 for the Long version. If this finding suggests a stronger stability than the cross-sectional versions, it must be underlined that a subtle bias could still go unnoticed due to the limited sample size.

3.3.3. Comparison between cross-sectional and longitudinal methods

In order to compare the proposed longitudinal version of BaMoS with the cross-sectional methods, the assessment measures were calculated across the 85 subjects for the mean assessment over each pattern. The corresponding results are summarised in Table 6. Due to the non-normality of the differences, a Wilcoxon test was applied to assess pairwise statistical significance between the DSC, TPR and AvDist. An increased performance in the order Cross < Cross+ < Long was observed and all tests were significant with p-value < 0.001 except for some comparisons between Cross+ and Long with the average distance that was only significant for the NonLinear_5 pattern and for the DSC in pattern Linear_750 that reached a *p*-value of 0.02. The comparison of the DSC across methods for the different progression patterns is presented in Fig. 4 illustrating the higher variances in assessment metrics for the cross-sectional methods compared to the proposed longitudinal version. Additionally, the robustness was tested through a linear regression of the ground truth volumes against the segmented volumes for all time points and subjects of the non-plateauing evolution patterns. Table 7 summarises the regression parameters, forcing the intercept to 0. In order to assess the bias with respect to the volume of lesions, Bland–Altman plots were drawn for all three methods. The difference between reference volume and segmented volume is plotted against the average of the two volumes in Fig. 5. The lines in the upper corners of the graphs for Cross and Cross+ correspond to cases with a very smooth appearance for which the cross-sectional methods had more difficulty in detecting lesions than Long. Conversely, the proposed method (Long) benefited from the increased signal-to-noise ratio of the group-mean to enable the detection of these subtle lesions.

Lastly, the longitudinal evolution of the DSC, corrected for the agreement volume at baseline, was compared across methods using a linear mixed model. Significant difference was observed in the slopes with a quicker decrease in DSC for the cross-sectional methods (p-value < 0.0001 and *p*-value =0.007 for Cross and Cross+ respectively). The corresponding plots of the predicted mean DSC are presented in Fig. 6. This illustrates, above the effect of the lesion volume, the impact of the smoothness of the lesions on the quality of the segmentation.

3.3.4. Simulation of treatment effect

For the 51 concerned cases, the slopes of white matter change were assessed using linear mixed models to compare the fast evolution (linear change of 750 mm^3 per step) and the combined

Table 3
Table summarising the slopes of volume change obtained for each simulated experiment for the three segmentation methods and the expected ground truth.

	Linear_500	Linear_750	NonLinear_5%	NonLinear_15
Cross	502.02 [422.14 581.9]	684.80 [525.88 843.73]	430.59 [347.97 513.21]	691.19 [527.99 854.38]
Cross+	436.81 [345.65 527.96]	670.58 [509.08 832.07]	297.01 [207.06 386.95]	832.45 [660.85 1004.05]
Long	277.21 [234.40 320.01]	550.02 [448.19 651.85]	168.96 [110.69 227.23]	638.55 [534.12 742.97]
Ref	424.96 [416.75 433.17]	703.49 [683.39 723.59]	247.08 [216.26 277.89]	861.56 [795.38 927.73]

Table 4
Segmentation assessment measures when evaluating the influence of plateauing stages on the longitudinal framework. By contrast to Flat_High and Flat_Low, Slope refers to a pattern without plateauing values. Results are given under the format median [IQR] and median are obtained across subjects on the average on common time points.

		DSC	TPR	AvDist	OE/TotF	OEFP/FP	OEFN/FN	FP/TotF
Linear_500	**Flat_Low**	0.68	0.83	1.89	0.81	0.70	0.96	0.78
		[0.27 0.76]	[0.62 0.89]	[1.14 11.91]	[0.46 0.89]	[0.41 0.83]	[0.85 0.98]	[0.62 0.86]
	Flat_High	0.67	0.84	1.88	0.81	0.73	0.96	0.81
		[0.28 0.76]	[0.68 0.90]	[1.21 10.63]	[0.51 0.89]	[0.40 0.86]	[0.86 0.98]	[0.67 0.89]
	Slope	0.65	0.83	2.07	0.81	0.72	0.97	0.83
		[0.28 0.77]	[0.66 0.91]	[1.00 9.83]	[0.49 0.90]	[0.41 0.87]	[0.88 1.00]	[0.60 0.92]
Linear_750	**Flat_Low**	0.60	0.77	2.92	0.76	0.68	0.93	0.81
		[0.36 0.73]	[0.59 0.88]	[1.19 9.96]	[0.57 0.87]	[0.49 0.84]	[0.75 0.98]	[0.66 0.88]
	Flat_High	0.67	0.78	2.10	0.76	0.71	0.94	0.81
		[0.34 0.78]	[0.61 0.88]	[0.95 9.91]	[0.56 0.90]	[0.50 0.86]	[0.76 0.97]	[0.62 0.87]
	Slope	0.66	0.80	1.89	0.77	0.71	0.93	0.79
		[0.34 0.76]	[0.57 0.88]	[1.06 9.69]	[0.53 0.88]	[0.47 0.85]	[0.78 0.97]	[0.59 0.86]
NonLinear_5	**Flat_Low**	0.65	0.81	1.98	0.81	0.73	0.96	0.76
		[0.41 0.76]	[0.67 0.89]	[1.08 5.38]	[0.68 0.89]	[0.55 0.84]	[0.91 0.98]	[0.60 0.83]
	Flat_High	0.68	0.79	1.99	0.82	0.71	0.96	0.75
		[0.41 0.76]	[0.66 0.88]	[1.18 5.65]	[0.66 0.89]	[0.56 0.85]	[0.90 0.98]	[0.56 0.85]
	Slope	0.66	0.81	1.93	0.82	0.74	0.97	0.77
		[0.41 0.76]	[0.65 0.88]	[0.96 5.46]	[0.67 0.90]	[0.55 0.87]	[0.92 1.00]	[0.56 0.87]
NonLinear_15	**Flat_Low**	0.64	0.80	1.95	0.82	0.70	0.96	0.79
		[0.37 0.73]	[0.71 0.88]	[1.17 6.71]	[0.60 0.88]	[0.51 0.85]	[0.92 0.98]	[0.69 0.85]
	Flat_High	0.65	0.80	1.81	0.8	0.71	0.94	0.73
		[0.38 0.76]	[0.66 0.86]	[1.06 6.51]	[0.61 0.89]	[0.50 0.85]	[0.90 0.97]	[0.58 0.84]
	Slope	0.64	0.79	2.05	0.8	0.69	0.94	0.76
		[0.41 0.74]	[0.67 0.85]	[1.11 6.62]	[0.57 0.88]	[0.50 0.85]	[0.90 0.97]	[0.59 0.82]

Acronyms expansion: DSC - Dice Similarity Coefficient; TPR - True Positive Rate; AvDist - Average Distance; FP/TotF - Proportion of false positives in the total of error; OE/TotF - Proportion of outline error in the total error; OEFP/FP - Proportion of false positive outline error in the false positives; OEFN/FN - Proportion of false negative outline error in the false negatives.

Table 5
Percentage of change between plateauing time points at either high or low volume and presented with median and IQR for each of the methods and the ground truth. The p-value corresponds to the Wilcoxon two-side test that the median is different from 0. For each type of plateau, the fourth row gives the Lin concordance between plateauing time points.

		Cross	Cross+	Long	Ref
Flat_High	**% change median**	3.14	2.28	0.21	0.09
	% change IQR	[−37.80 50.00]	[−17.77 25.11]	[−8.19 9.29]	[−1.99 1.89]
	***p*-value**	0.02	0.06	0.54	0.83
	Lin concordance	0.81	0.95	0.98	0.999
Flat_Low	**% change median**	−3.52	6.53	1.53	0
	% change IQR	[−52.9 111.3]	[−17.1 7.78]	[12.0 17.6]	[−1.96 1.46]
	***p*-value**	0.0021	0.0002	0.05	0.42
	Lin concordance	0.49	0.87	0.95	0.999

evolution (initial step of 750 ^{3} change followed by two steps of 500 ^{3} change) after bifurcation of the evolutions. The results of the estimation and confidence interval are presented in Table 8 illustrating notably the lower measurement variance observed for the longitudinal framework compared to the cross-sectional methods. Noticeably, the LST method with threshold 0.25 did not appear to segment properly the lesions.

3.3.5. Simulation of lesion growing with atrophy – Robustness to order

When assessing the robustness of the longitudinal method to the order in which images were considered to build the average, the linear regression between volumes obtained in the forward or the backward scheme led to a R2 of 0.99 and the median DSC between the two segmentations was of 0.98. Statistics on the difference in DSC (wrt the ground truth) per time point are gathered in Table 9. With respect to other cross-sectional methods, Table 10 presents the evolution of DSC with atrophy across methods while Fig. 7 compares the different segmentations for a given simulated sequence.

4. Application to clinical data

4.1. Data and experiment

Although no ground truth is available for the lesion segmentation in clinical practice, a surrogate validation consists of testing it against known clinical findings. In the case of the proposed longitudinal framework, both cross-sectional and longitudinal observations can be tested. Here, the longitudinal framework was applied on the subjects from the ADNI (Alzheimer's Disease National Initiative) database (adni.loni.usc.edu) for which T1 and FLAIR images on at least four time points were available along with genetic status for *APOE*, that in its allelic variant $\epsilon4$ (among $\epsilon2$ $\epsilon3$ and $\epsilon4$) is a known risk factor for AD and a presumed risk factor for white matter lesions. The *APOE* status refers to the two alleles of the gene present in a given individual. Due to its very low prevalence in the population (about 2%), the *APOE*-$\epsilon2$ were discarded from the analysis, thus allowing for the combination $\varepsilon3\varepsilon3$, $\varepsilon4\varepsilon3$ or $\varepsilon4\varepsilon4$. Launched in 2003, ADNI's primary goal is to test whether the combination of serial MRI and other biological and neuropsychological markers is relevant to assess the development

Table 6
Segmentation assessment comparison for the mean over time points of the three compared methods across all subjects for all non plateauing patterns. Results are given under the form median [IQR].

		DSC	TPR	AvDist	OE/F	OEFP/FP	OEFN/FN	FP/F
Linear_500	**Cross**	0.26 [0.09 0.39]	0.21 [0.09 0.33]	14.54 [6.97 21.42]	0.52 [0.28 0.70]	0.39 [0.25 0.57]	0.62 [0.37 0.75]	0.13 [0.04 0.33]
	Cross+	0.58 [0.26 0.74]	0.64 [0.43 0.83]	5.59 [1.17 13.66]	0.75 [0.45 0.89]	0.68 [0.41 0.85]	0.89 [0.67 0.96]	0.64 [0.49 0.74]
	Long	0.65 [0.28 0.77]	0.83 [0.66 0.91]	2.07 [1.00 9.83]	0.81 [0.49 0.90]	0.72 [0.41 0.87]	0.97 [0.88 1.00]	0.83 [0.60 0.92]
Linear_750	**Cross**	0.28 [0.12 0.46]	0.20 [0.09 0.35]	11.45 [3.51 22.29]	0.56 [0.37 0.79]	0.46 [0.31 0.61]	0.62 [0.34 0.82]	0.11 [0.03 0.28]
	Cross+	0.64 [0.30 0.75]	0.69 [0.39 0.81]	2.06 [1.02 11.02]	0.79 [0.53 0.88]	0.70 [0.42 0.85]	0.84 [0.59 0.93]	0.60 [0.42 0.70]
	Long	0.66 [0.34 0.76]	0.80 [0.57 0.88]	1.89 [1.06 9.69]	0.77 [0.53 0.88]	0.71 [0.47 0.85]	0.93 [0.78 0.97]	0.79 [0.59 0.86]
NonLinear_5	**Cross**	0.21 [0.09 0.34]	0.16 [0.07 0.27]	14.03 [8.22 22.49]	0.51 [0.3 0.68]	0.35 [0.24 0.50]	0.54 [0.28 0.72]	0.07 [0.04 0.16]
	Cross+	0.56 [0.33 0.75]	0.69 [0.41 0.81]	3.47 [1.12 9.03]	0.78 [0.57 0.89]	0.65 [0.43 0.85]	0.88 [0.66 0.95]	0.58 [0.36 0.69]
	Long	0.66 [0.41 0.76]	0.81 [0.65 0.88]	1.93 [0.96 5.46]	0.82 [0.67 0.90]	0.74 [0.55 0.87]	0.97 [0.92 1.00]	0.77 [0.56 0.87]
NonLinear_15	**Cross**	0.25 [0.11 0.38]	0.19 [0.07 0.28]	11.70 [6.02 17.65]	0.53 [0.37 0.70]	0.42 [0.31 0.60]	0.55 [0.36 0.76]	0.10 [0.03 0.20]
	Cross+	0.57 [0.31 0.73]	0.65 [0.43 0.78]	2.64 [1.11 7.21]	0.79 [0.56 0.89]	0.64 [0.45 0.82]	0.89 [0.68 0.94]	0.55 [0.37 0.71]
	Long	0.64 [0.41 0.74]	0.79 [0.67 0.85]	2.05 [1.11 6.62]	0.8 [0.57 0.88]	0.69 [0.50 0.85]	0.94 [0.90 0.97]	0.76 [0.59 0.82]

Acronyms expansion: DSC - Dice Similarity Coefficient; TPR - True Positive Rate; AvDist - Average Distance; FP/TotF - Proportion of false positives in the total of error; OE/TotF - Proportion of outline error in the total error; OEFP/FP - Proportion of false positive outline error in the false positives; OEFN/FN - Proportion of false negative outline error in the false negatives.

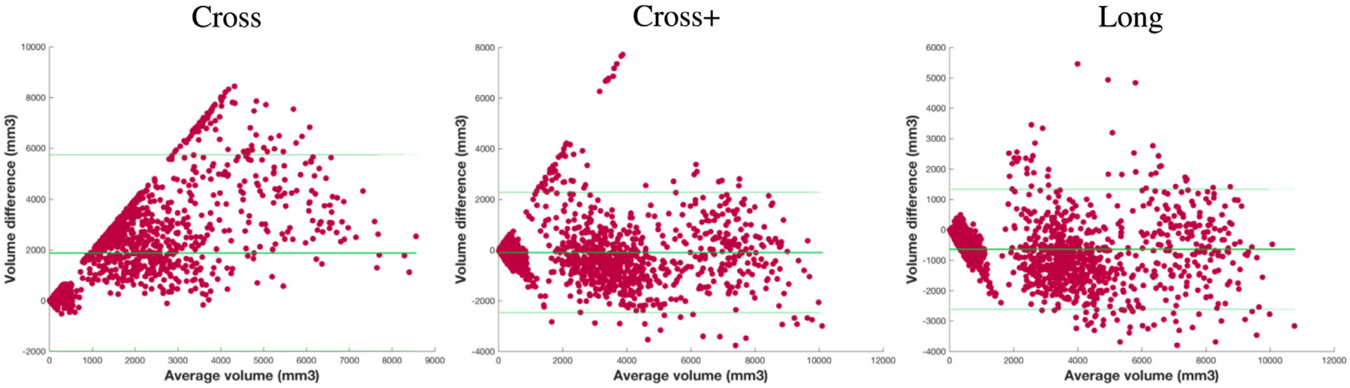

Fig. 5. Bland–Altman plots of the reference and segmented volumes.

Table 7
Coefficients and corresponding 95% confidence intervals of the regression between segmented and reference volume for the three compared segmentation methods. Regression is performed to account for the within subject correlations in volumes.

	Cross	Cross+	Long
Slope	0.31 [0.26 0.36]	0.98 [0.90 1.06]	1.12 [1.05 1.20]
R^2	0.60	0.90	0.94

Table 9
Statistics of the difference in DSC when considering the forward or backward order when building the average image.

	Mean	SD	Median	IQR
TP1	−0.0036	0.027	−0.0009	[−0.0072 0.0055]
TP2	−0.0015	0.018	−0.0002	[−0.0100 0.0034]
TP3	−0.0007	0.013	0.0006	[−0.0037 0.0031]
Total	−0.0019	0.020	−0.0003	[−0.0046 0.0034]

Table 8
Slope estimation after evolution bifurcation for the three methods. Mean and confidence intervals (CI) for the two slopes are given.

		Cross	Cross+	Long	LPA	LST-d	LST-a	TODAS
Treatment	**Mean**	439	532	370	316	−4.4	0.47	474
	CI	[282 596]	[383 681]	[269 471]	[34 598]	[−13.7 4.9]	[−10.9 11.8]	[−684 1632]
No treatment	**Mean**	786	782	627	332	−8.7	−9.93	1559
	CI	[597 975]	[629 935]	[525 729]	[50 613]	[−18.0 0.6]	[−21.3 1.4]	[400 2717]
Statistics	***p*-value**	0.006	0.022	0.0003	0.93	0.48	0.20	0.18

Table 10
DSC for each time point after baseline for the different methods compared.

		Cross	Cross+	Long	LPA	LST-d	LST-a	TOADS
TP 1	Median	0.21	0.67	0.68	0.65	0	0	0.63
	IQR	[0.01 0.47]	[0.61 0.75]	[0.62 0.75]	[0.56 0.74]	[0 0.004]	[0 0.008]	[0.52 0.73]
TP 2	Median	0.10	0.73	0.72	0.69	0	0	0.67
	IQR	[0.01 0.45]	[0.66 0.77]	[0.66 0.77]	[0.61 0.76]	[0 0.0008]	[0 0.005]	[0.49 0.74]
TP 3	Median	0.46	0.68	0.68	0.71	0.003	0.004	0.73
	IQR	[0.16 0.57]	[0.64 0.72]	[0.64 0.73]	[0.64 0.77]	[0 0.01]	[0 0.02]	[0.66 0.80]
Global	Median	0.32	0.69	0.69	0.70	0	0.002	0.69
	IQR	[0.02 0.51]	[0.64 0.75]	[0.64 0.75]	[0.61 0.76]	[0 0.005]	[0 0.008]	[0.55 0.75]

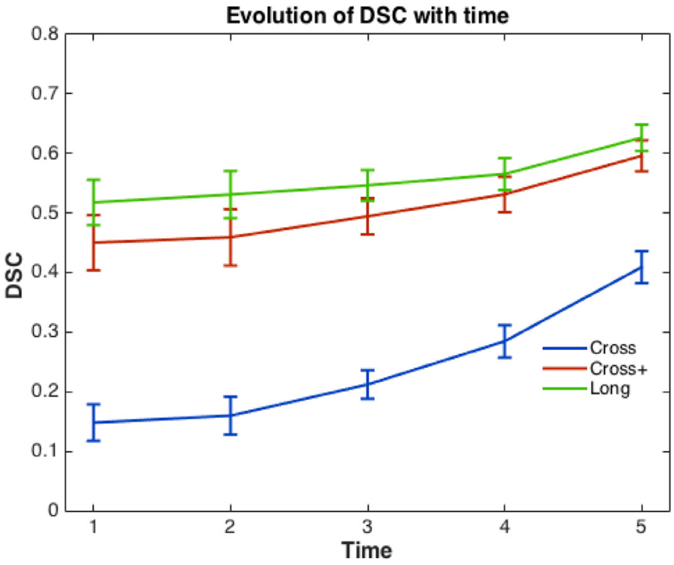

Fig. 6. Mean DSC evolution with 95% CI when correcting for baseline volume across the three evaluated methods.

of Alzheimer's disease and the progression of mild cognitive impairment. Further information about imaging and genetic protocols can be found at www.adni-info.org. Focusing here on WMH, the effect of age as risk factor was assessed on the baseline volumes.

Since in ADNI, some subjects undergo MRI scanning sessions at a small time interval ($<$ 5 months), during which, the volume of WMH is assumed to change subtly and the variability of change across subjects is supposed to be reduced, such series of scans were used as a surrogate for a short term change experiment to evaluate the ability of various methods to detect subtle change and the variability in the measured changes. Therefore, both the monthly volume difference and relative difference between close time points were compared across methods.

Additionally, in ageing, the amount of WMH is not expected to decrease in time. Thus a comparison of spurious decrease between methods is a further way of evaluating the clinical potential of the methods evaluated. In order to assess the consistency between time points, the mean ratio of decrease for each subject per month was compared across methods. Although some noise in the measures may be expected, smoother volume trajectories are thought to reflect more consistency in the segmentation results.

Then, in order to investigate the relationship between WMH accumulation and *APOE* genetic status, generalised linear mixed models enabling the analysis of repeated measures were applied. Due to the skewness of their distribution, WMH volumes were modelled as following a gamma distribution and used as the dependent variable while time from initial measurement was considered both as fixed and random effects thus allowing for individual slopes and intersection values. The other fixed effects included age, sex, total intracranial volume and *APOE* status, and their interaction with time. After adjustment for covariates, joint Wald tests were used to compare rates of change between *APOE* status. The presented results are the fitted mean rates of change, standardised to the mean levels of covariates in the sample as a whole. Eleven cases failed to provide segmentation results for LST while two failed for LPA. To ensure model convergence, priors were imposed on the covariance matrices that may otherwise be naturally singular.

4.2. Results

The demographics of the subjects are gathered in Table 11 and the measurements obtained at the first time point are reported for the different segmentation methods used. TOADS appeared to strongly oversegment lesions in some cases. When correcting for age, sex and TIV, the age factor was found to be strongly positively associated with the white matter volume at the earliest time point ($p < 10^{-5}$). The raw relationship between age and WMH volume is displayed in Fig. 8.

With respect to the measures of repeatability for time points with a short interval, 274 subjects satisfied the criteria of an interval of 4 months or less between the first two measurements. Table 12 summarises both mean and relative differences when comparing the different segmentation methods. Wilcoxon pair signed tests highlighted mean difference significantly different from 0 for all tested methods (p-value < 0.005 all tests) except TOADS ($p = 0.68$) with a smaller variability for Long compared to other cross-sectional strategies as illustrated in Fig. 9. Then, assessments evaluating the mean spurious absolute and relative decrease in the time series, showed also significantly less inadequate WMH volume decrease in Long compared to the cross-sectional methods. Quantitative results are gathered in Table 13. Although the difference was borderline significant when comparing Long and LST-a, the distribution appeared again more concentrated for Long, thereby underlining the robustness of the result as shown in Fig. 10. An example of segmentations for the different methods is illustrated in Fig. 11.

The results for the longitudinal analysis performed for both longitudinal and cross-sectional methods are summarised in Table 14 presenting the mean rate of change in WMH volume per year. Using both methods, the rate of change increased with the number of *APOE* $\varepsilon 4$ alleles.

To better illustrate the differences between the methods, Table 15 summarises the pairwise effect size when comparing the slopes of WMH progression across genetic status groups.

5. Discussion

In this work the methodological aspects of a new longitudinal framework for WMH segmentation accounting for within-subject consistency was detailed and tested with synthetic and clinical data. The validation on synthetic data was enabled by the use of

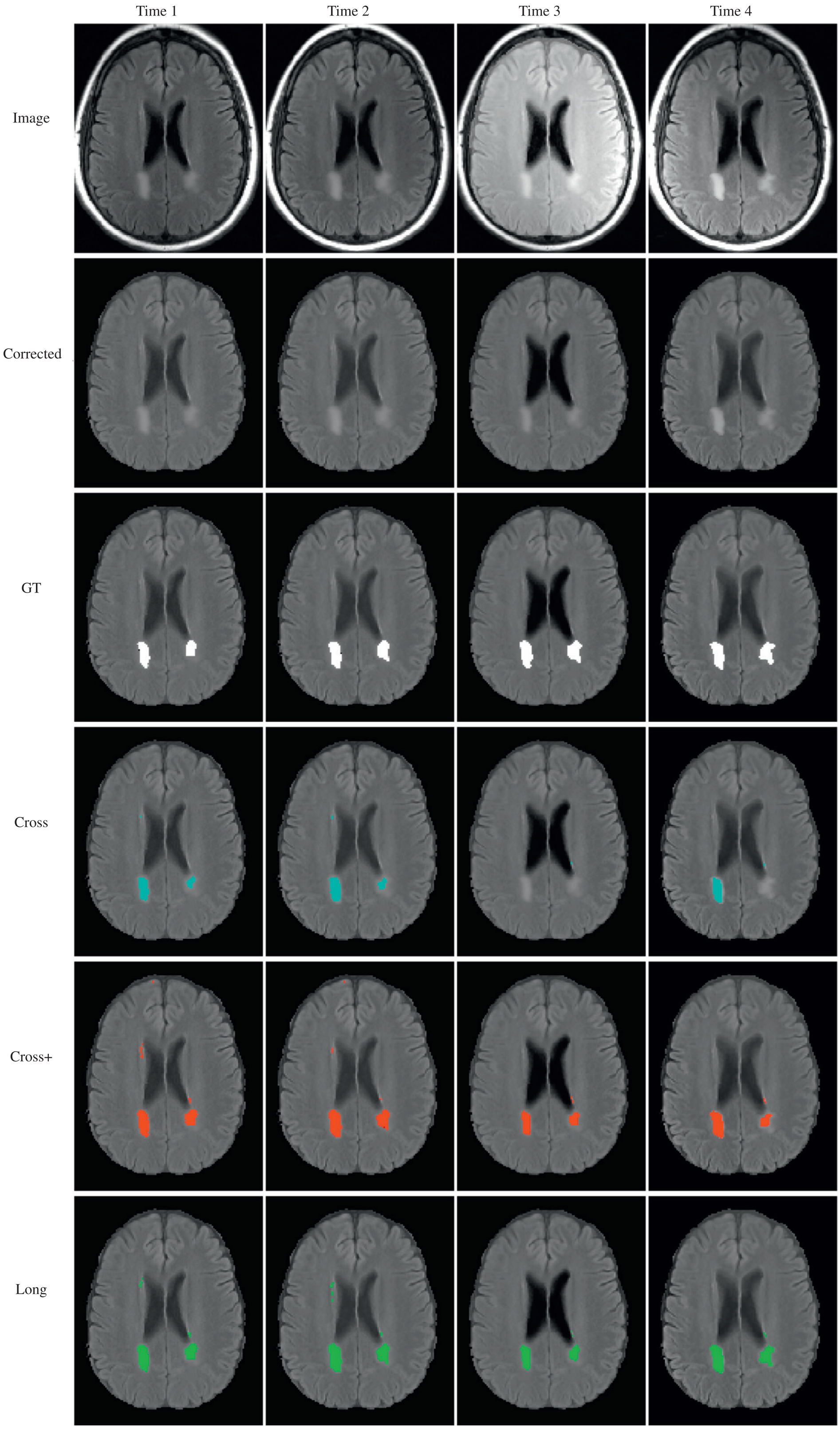

Fig. 7. Example of simulated images with atrophy pattern and an increase of 750 mm^3 per step. The ground truth (GT, third row) and compared segmentation results (row 4 to 6) are overlayed on the corrected (2nd row) FLAIR images. These corrected images display the log-transformed normalised, bias field corrected, skull-stripped and intensity matched intensities.

Table 11
Demographics of the studied ADNI population at baseline.

		APOE			Global
		33	43	44	
Number Tot [Female]		164 [76]	108 [52]	24 [6]	300 [137]
Age mean (SD)		72.9 (7.0)	71.3 (7.7)	71.1 (7.32)	72.1 (7.3)
TIV (mL) mean (SD)		1552 (143)	1549 (171)	1564 (141)	155.2 (15.3)
Time span (year) mean (SD)		2.01 (0.74)	2.01 (0.86)	1.84 (0.66)	2.00 (0.78)
WMH (mL) median [IQR]	**Cross**	1.21 [0.58 3.33]	1.37 [0.50 3.52]	1.65 [0.65 3.71]	1.23 [0.56 3.39]
	Cross+	1.82 [0.96 5.48]	2.44 [0.81 5.90]	2.74 [1.17 5.59]	1.96 [0.94 5.54]
	Long	2.06 [0.88 5.42]	2.48 [0.68 6.08]	2.93 [0.96 5.64]	2.15 [0.96 5.64]
	LPA	2.31 [1.03 7.33]	2.88 [0.86 8.49]	3.70 [1.50 8.75]	2.65 [0.96 7.96]
	LST-d	1.64 [0.48 6.42]	2.22 [0.61 5.74]	2.75 [0.84 5.34]	1.94 [0.58 5.56]
	LST-a	1.91 [0.69 6.86]	2.64 [0.75 6.16]	3.12 [1.05 5.84]	3.05 [0.71 6.22]
	TOADS	24.32 [9.35 44.27]	22.44 [10.45 52.75]	29.63 [14.28 57.86]	24.10 [10.15 48.69]

Acronyms expansion: TIV - Total Intracranial Volume; WMH - White Matter Hyperintensities; IQR - InterQuartile Range; SD - standard deviation.

Table 12
Measures of variation of WMH volume segmentation for scan sessions with less than 4 months interval. The relative difference is the ratio of the difference divided by the volume at the second time point.

	Method	Median	IQR	Mean	SD	Range
Percentage change	**Cross**	1.40	[−3.24 5.02]	−0.01	9.65	[−46.33 24.57]
	Cross+	0.68	[−2.94 4.26]	4.26	14.43	[−115.70 21.64]
	Long	0.67	[−2.49 2.65]	2.65	8.09	[−50.19 19.59]
	LPA	1.46	[−1.73 4.00]	4.00	93.24	[−1533.62 29.13]
	LST-d	0.76	[−2.85 4.24]	−2.85	18.78	[−219.89 33.33]
	LST-a	0.48	[−2.31 3.61]	−2.31	19.88	[−266.67 33.33]
	TOADS	0.45	[−6.71 4.79]	−18.14	92.24	[−908.20 32.13]
Difference	**Cross**	22.41	[−34.58 113.46]	68.35	347.85	[−1390.91 3441.22]
	Cross+	23.93	[−50.47 126.13]	58.70	338.26	[−1632.72 3242.80]
	Long	13.215	[−33.99 99.91]	38.93	232.43	[−947.80 2382.90]
	LPA	29.09	[−29.60 136.79]	22.91	1706.64	[−26718.34 5286.04]
	LST-d	9.51	[−34.693 87.83]	42.34	347.97	[−2396.03 2620.60]
	LST-a	8.51	[−41.45 86.01]	42.13	368.9537	[−2660.41 2763.80]
	TOADS	107.14	[−1171.10 1100.33]	−166.49	6203.99	[−50877.84 27942.62]

Table 13
Average relative and absolute spurious decrease per subject across the different compared methods.

	Method	Median	IQR	Mean	SD	Range
Percentage change	**Cross**	2.99	[1.06 6.89]	5.72	8.98	[0 93.76]
	Cross+	2.76	[0.92 7.68]	10.31	40.65	[0 612.49]
	Long	1.74	[0.58 4.02]	4.22	12.78	[0 201.59]
	LPA	1.51	[0.22 3.60]	743.59	9583.02	[0 150815.30]
	LST-d	1.96	[0.49 4.99]	6.53	20.73	[0 273.83]
	LST-a	1.96	[0.40 4.96]	6.21	17.24	[0 165.80]
	TOADS	4.82	[1.66 14.91]	29.02	83.35	[0 770.33]
Mean decrease	**Cross**	37.48	[15.27 87.43]	83.40	145.15	[0 1291.01]
	Cross+	56.42	[19.91 126.58]	103.24	137.01	[0 821.59]
	Long	31.23	[9.01 75.81]	67.64	107.71	[0 858.6]
	LPA	30.24	[2.15 100.53]	174.46	857.86	[0 13391.54]
	LST-d	30.05	[4.57 93.30]	87.12	164.53	[0 1525.26]
	LST-a	31.32	[7.24 93.06]	93.61	180.82	[0 1574.36]
	TOADS	1023.43	[254.74 2779.38]	2245.22	4013.40	[0 50877.84]

Table 14
Longitudinal analysis of WMH volumes across genetic *APOE* status obtained with the proposed longitudinal method and the other cross-sectional solutions. Mean rate of volume change per year are presented with their confidence intervals (CI) when adjusting for age, sex and total intracranial volume. Slopes are given in mL/year.

	33		43		44		*p*-values
	Mean	CI	Mean	CI	Mean	CI	
Cross	0.19	[0.13 0.25]	0.36	[0.26 0.46]	1.24	[0.66 1.82]	33 vs 43 0.008 33 vs 44 0.0005 43 vs 44 0.0005
Cross+	0.35	[0.23 0.47]	0.41	[0.24 0.58]	1.02	[0.37 1.66]	33 vs 43 0.993 33 vs 44 0.061 43 vs 44 0.066
Long	0.17	[0.12 0.22]	0.30	[0.21 0.39]	0.91	[0.46 1.35]	33 vs 43 0.026 33 vs 44 0.0005 43 vs 44 0.0005
LPA	0.50	[0.37 0.62]	1.06	[0.80 1.32]	2.36	[1.23 3.49]	33 vs 430.0005 33 vs 44 0.0005 43 vs 44 0.011
LST-d	0.38	[0.28 0.47]	0.59	[0.43 0.74]	1.28	[0.62 1.94]	33 vs 430.075 33 vs 44 0.005 43 vs 44 0.065
LST-a	0.40	[0.30 0.50]	0.66	[0.49 0.82]	1.50	[0.75 2.26]	33 vs 430.025 33 vs 44 0.001 43 vs 44 0.027
TOADS	1.08	[−0.51 2.66]	1.09	[−0.83 3.00]	2.10	[−2.97 7.18]	33 vs 430.993 33 vs 44 0.738 43 vs 44 0.740

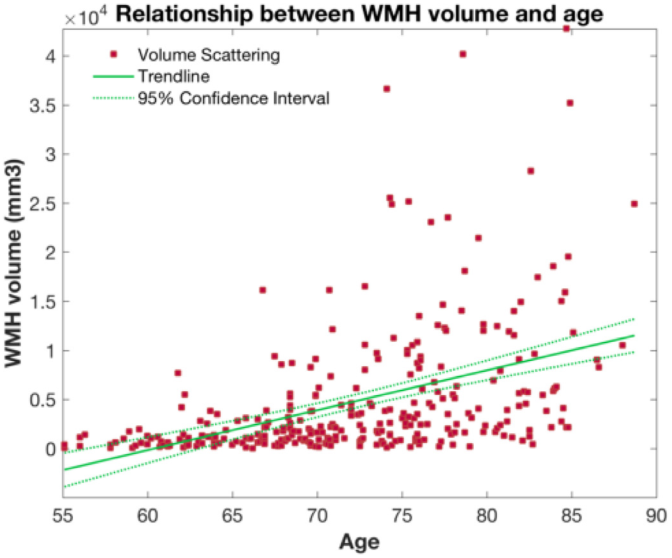

Fig. 8. Relationship between age and WMH both taken at baseline.

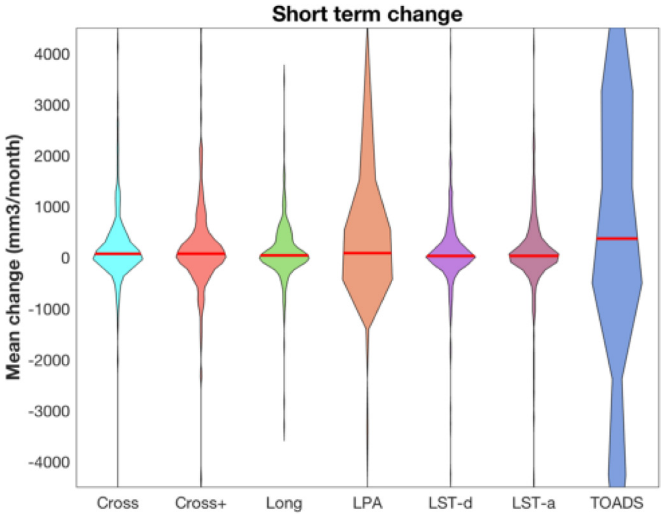

Fig. 9. Density plot of the variation observed at low time interval in terms of mean change per month across segmentation methods.

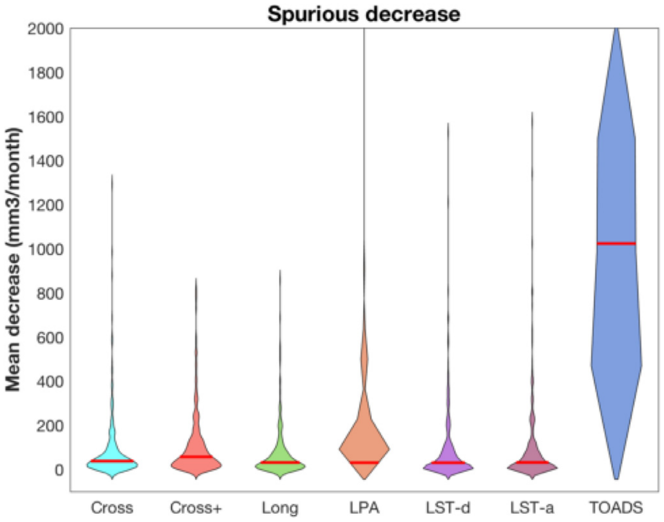

Fig. 10. Density plot of the average decrease observed in terms of percentage of change per month across segmentation methods.

Table 15
Effect sizes when comparing slopes of WMH progression in the different methods across genetic APOE status.

	33 vs 43	43 vs 44	33 vs 44
Cross	0.24	0.36	0.52
Cross+	0.05	0.21	0.28
Long	0.21	0.32	0.47
LPA	0.34	0.26	0.47
LST-d	0.20	0.24	0.39
LST-a	0.22	0.26	0.42
TOADS	0	0.04	0.05

a lesion simulator incorporating WMH lesion evolution patterns. A major difficulty in the assessment of segmentation protocols and algorithms is indeed the availability of a possible ground truth. The use of simulated data as an initial step has been promoted in Lladó et al. (2012) for lesion segmentation. Actually, simulation of white matter lesions at different loads has been made available in the Brainweb project http://brainweb.bic.mni.mcgill.ca. However no simulated longitudinal progression is available. Here, the simulation is applied to existing clinical data so that the realism of the simulated images is high. It must however be noted that the lesion simulator, although tested with different lesion loads, is based on typical age-related lesion distribution patterns as observed in the ADNI dataset. Increasing the variability of the lesion maps as well as the anatomical shape of change used for the simulator would be of further interest. Furthermore, extensions towards MS-like patterns of evolution, involving the explicit modelling of appearance and disappearance of lesions could be added, using for instance secondary lesion maps in the simulation.

With respect to the longitudinal framework, this simulator allowed the evaluation of the impact of evolution patterns on the segmentation performance as well as the impact of periods without change on the progression detection. From the assessment of the longitudinal framework across evolution patterns, subjects with smaller variations in WMH load led to slightly better overall segmentation results although no statistical difference was observed. Investigating longitudinal bias through plateauing situations, the stability in the results highlights the ability of the longitudinal framework to detect change even if periods of stability are included. This stability is crucial in processes for which subtle and irregular progressions are observed, such as multiple sclerosis. In terms of error, most of the erroneous classifications appear at the border of the lesions and very few lesions were completely undetected. The comparison between the longitudinal and the cross-sectional versions of BaMoS (Cross and Cross+) underlines the improved robustness of the proposed framework with higher performance and lower variance in the results. Although the detected volumes appeared higher than expected, the strong correlation observed between segmented and reference volumes (R^2=0.94) makes the detection of change trustworthy. This tendency to overestimation can be partially related to the observation, that the longitudinal framework tended to underestimate the true slope. It can be explained by a bias of the model towards the average image with less noise and therefore smaller covariance matrices that in turn contribute to a less conservative outlier separation.

The ability to detect differences in longitudinal rate of change was however exemplified in the simulations of treatment effect. In this case, the difference observed between evolutions is similar to the simulated ground truth difference. A decreased variance reflecting higher measurement robustness compared to both cross-sectional methods led to a decrease in required sample size. In this setting however, the LST algorithm appeared to perform poorly. Using an experiment with atrophy simulation, thus incorporating

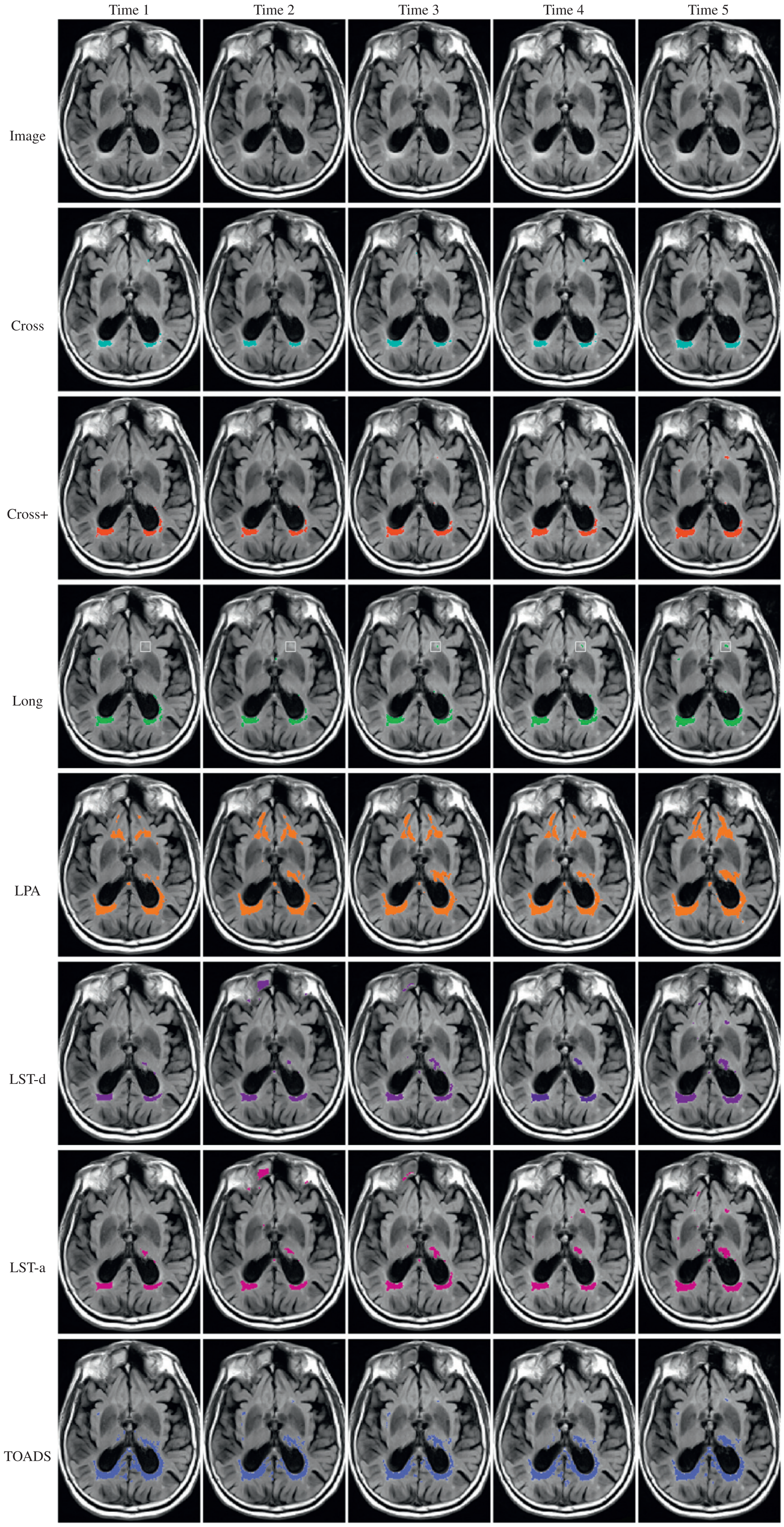

Fig. 11. Example of resulting segmentations over time for the five evaluated methods. For visualisation purposes, segmentations have been registered to the groupwise space and binarised. The first row presents the FLAIR images and the subsequent rows the overlayed segmentation results. For the longitudinal segmentation, the white boxes highlight the monotonous increase in volume for a specific lesion.

stronger volume change, the susceptibility of bias towards image ordering was investigated. In the proposed framework, the only source of potential bias lies in the order used to build the average image on which the model used to constrain the segmentation of the individual time points is selected. The iterative process used to build the average image is assumed to prevent bias to any specific time point. Comparing segmentation results obtained using the forward (increasing lesion load and atrophy) and backward (decreasing lesion load and atrophy) orders showed that the results were stable with respect to the order with a R^2 between segmented volumes of 0.96 and a median difference DSC of less than 0.03% of DSC at each evaluated time point. This agnosticity to order and therefore independence with respect to lesion evolution would make this algorithm suitable for an extension to the longitudinal analysis of multiple sclerosis patients. In this setting, the longitudinal method appeared to be more robust for dealing with global volume changes as shown by a lower variability in DSC.

When applying the longitudinal framework on clinical data, in the case of the ADNI population, findings reported in the literature both cross-sectionnally and longitudinally were reproduced using the proposed method. As such, age, established risk factor for the existence of WMH (Grueter and Schulz, 2012; Targosz-Gajniak et al., 2009; Schmahmann et al., 2008), was strongly associated with the segmented volumes. In a short term change experiment using the first two time points when spaced by less than 5 months, the mean difference detected was for all methods except TOADS significantly different from 0 but the variability lower for the longitudinal framework. Therefore, the worries that Long may miss some changes due to an underestimation bias are alleviated by the consistency it shows in its detection of change.

As far as the longitudinal analysis of rate of change was concerned, a dose-dependent effect of *APOE* $\varepsilon 4$ was observed for which homozygous $\varepsilon 4 \varepsilon 4$ were found to progress faster than heterozygous and non-carriers with regards to WMH load which has been reported in the literature (Godin et al., 2009). Noticeably, the effect size of this finding was stronger when using a longitudinal method compared to a cross-sectional analysis. The differences observed may be related to the increased variability introduced when considering data cross-sectionally. Note however that the rates of change were significantly different between methods; in particular, LST was found to strongly overestimate volumes. However, observed discrepancies in absolute values between methods can be mitigated by studying the relative change between groups and estimating statistical group differences as commonly done in clinical trials.

To conclude, a new longitudinal framework for WMH segmentation was presented and an increased robustness was demonstrated compared to similar methods applied cross-sectionally, thereby ensuring the relevance of the longitudinal extension of the original method. For validation purposes, a realistic longitudinal lesion simulator was developed allowing for a wide variety of evolution patterns on a possibly large range of images. Assessments relative to the images order in the building of the average image showed a high consistency thus ensuring the absence of bias with respect to the ordering of images. Clinically, when applied to ADNI, a large cohort of elderly with minimal cardiovascular risk factors, previously reported cross-sectional and longitudinal findings were again noticed and the longitudinal method proved more powerful in highlighting group differences than the cross-sectional methods it was compared to. Further work could include the evaluation of the impact of the parameters chosen to build the average appearance model. Future research will also work towards reducing the between-time-point group-average transformation regularisation so as to maximise not only the precision but also the accuracy of the measurements.

Acknowledgements

Carole H. Sudre is funded by the Wolfson Foundation and the UCL Faculty of Engineering M. Jorge Cardoso receives funding from EPSRC (EP/H046410/1). Sebastien Ourselin receives funding from the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH -DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative- BW.mn.BRC10269).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81 XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix

Description of measures of segmentation evaluation. When comparing a segmentation result (Seg) to a segmentation reference (Ref) used as gold standard or ground truth, different assessments made either at the voxel level or at the cardinal/entity level can be performed. In the case of lesion segmentation, an entity corresponds to a set of connected voxels segmented as lesion. False positives (FP) are defined as the elements (either voxel and entity) detected in Seg but not present in Ref while false negatives (FN) are present in Ref but omitted in Seg. True positives are the elements detected in Seg and truly present in Ref. At the cardinal level a lesion in Seg that has at least one voxel in Ref is considered as a true positive. The cardinal definitions are assigned the subscript c. Using sets definition with $\sharp$ referring to the number of elements of a set, $FP = \sharp Seg \bigcap R\bar{e}f$, $FN = \sharp S\bar{e}g \bigcap Ref$ and $TP = \sharp Seg \bigcap Ref$.

DSC Classical measure of overlap evaluation, the Dice score coefficient (DSC) is expressed as $DSC = \frac{2TP}{2TP + FN + FP} = \frac{2\sharp Seg \bigcap Ref}{\sharp Seg + \sharp Ref}$

AvDist The average distance, mentioned by Datta and Narayana (2013) and Styner et al. (2008) measures the average distance between the two lesion outlines

$$\mathbf{AvDist}(\mathrm{Ref}, \mathrm{Seg}) = \frac{\sum_{s \in \partial \mathrm{Seg}} \min_{s \in \partial \mathrm{Seg}} d(s, r) + \sum_{r \in \partial \mathrm{Ref}} \min_{s \in \partial \mathrm{Seg}} d(s, r)}{\sharp_v \partial \mathrm{Seg} + \sharp_v \partial \mathrm{Ref}}$$

where ∂Seg (resp. ∂*Ref*) denotes the border in the 18-neighbour connectivity of the Seg (resp. Ref) set. and $d(s, r)$ is the Euclidean distance between element s and r. It must be however noted the difficult definition of such an assessment measure when one of the volumes is 0.

True positive rate (TPR) The true positive rate (TPR) that can be defined either at the voxel or cardinal level is expressed as $\frac{\sharp \mathrm{TP}}{\sharp \mathrm{Ref}}$ and takes its values in [0 ; 1] with 1 as best value. With this measure, a perfect score at the voxel level can be reached for a suboptimal segmentation if the errors are exclusively false positives. In the cardinal form, the ratio becomes dependent of the lesion spatial connectivity since joined lesion are only counted once.

Recently, new inter-rater assessment measures with application to MS lesion segmentation have been developed. These have shown to be less dependent than the DSC to the assessed lesion burden (Wack et al., 2012).

Detection error (DE) The detection error is the volume of error measured cardinally (using lesion entities) and is expressed as

$$\mathrm{DE} = \sum_{F \in \mathrm{FP}_c} \sharp_v \mathrm{Seg}_F + \sum_{F \in \mathrm{FN}_c} \sharp_v \mathrm{Ref}_F .$$

Outline error (OE) The OE is measured as the volume of voxelwise error found for the true positive lesion entities.

$$\mathrm{OE} = \sum_{T \in \mathrm{TP}_c} \sharp_v (\mathrm{Seg}_T \cup \mathrm{Ref}_T) - \sharp_v (\mathrm{Seg}_T \cap \mathrm{Ref}_T).$$

In order to better assess the origin of the errors, based on the definition by Wack et al. (2012), additional evaluation may be carried out:

OE/F Measures the proportion of total error (F) that is related to the outline error OE.

FP/F Measures the proportion of error that is false positive.

OEFP/FP Measures among the false positives, the proportion that relates to the outline error.

OEFN/FN Measures among the false negatives, the proportion that relates to the outline error.

References

Abraham, H.M.A., Wolfson, L., Moscufo, N., Guttmann, C.R.G., Kaplan, R.F., White, W.B., 2015. Cardiovascular risk factors and small vessel disease of the brain: blood pressure, white matter lesions, and functional decline in older persons. J. Cereb. Blood Flow Metab. (April) 1–7. doi:10.1038/jcbfm.2015.121.

Bosc, M., Heitz, F., Armspach, J.-P., Namer, I., Gounot, D., Rumbach, L., 2003. Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. Neuroimage 20, 643–656. doi:10.1016/S1053-8119(03)00406-3.

Cardoso, M.J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., Ourselin, S., 2015. Geodesic information flows: spatially-Variant graphs and their application to segmentation and fusion. IEEE Trans. Med. Imaging 34 (9), 1976–1988. doi:10.1109/TMI.2015.2418298.

Carmichael, O., Schwarz, C., Drucker, D., Fletcher, E., Harvey, D., Beckett, L., Jack, C.R., Weiner, M., DeCarli, C., 2010. Longitudinal changes in white matter disease and cognition in the first year of the alzheimer disease neuroimaging initiative.. Arch. Neurol. 67 (11). doi:10.1001/archneurol.2010.284. 1370–8

Datta, S., Narayana, P.A., 2013. A comprehensive approach to the segmentation of multichannel three-dimensional MR brain images in multiple sclerosis. NeuroImage 2, 184–196.

Elliott, C., Arnold, D.L., Collins, D.L., Arbel, T., 2013. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. IEEE Trans. Med. Imaging 32 (8), 1490–1503. doi:10.1109/TMI.2013.2258403.

Godin, O., Tzourio, C., Maillard, P., Alpérovitch, A., Mazoyer, B., Dufouil, C., 2009. Apolipoprotein e genotype is related to progression of white matter lesion load. Stroke 40 (10), 3186–3190. doi:10.1161/STROKEAHA.109.555839.

Gouw, A.a., Van Der Flier, W.M., van Straaten, E.C.W., Pantoni, L., Bastos-Leite, A.J., Inzitari, D., Erkinjuntti, T., Wahlund, L.-O.O., Ryberg, C., Schmidt, R., Fazekas, F., Scheltens, P., Barkhof, F., 2008. Reliability and sensitivity of visual scales versus volumetry for evaluating white matter hyperintensity progression. Cerebrovascular Dis. 25, 247–253. doi:10.1159/000113863.

Grueter, B.E., Schulz, U.G., 2012. Age-related cerebral white matter disease (leukoaraiosis): a review. Postgrad. Med. J. 88, 79–87.

Gunning-Dixon, F.M., Raz, N., 2000. The cognitive correlates of white matter abnormalities in normal aging: a quantitative review. Neuropsychology 14 (2), 224–232. doi:10.1037/0894-4105.14.2.224.

Jack Jr, C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Yuecheng, X., Muthupillai, R., Manduca, A., Avula, R., Erickson, B.J., Jack, C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., Erickson, B.J., 2001. FLAIR Histogram segmentation for measurement of leukoaraiosis volume. J. Magn. Reson. Imaging 14, 668–676. doi:10.1002/jmri.10011.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrentà, L., Rovira, À., Ramió-Torrent, L., Rovira, L., 2012. Segmentation of multiple sclerosis lesions in brain MRI : a review of automated approaches. Inf. Sci. (Ny) 186, 164–185. doi:10.1016/j.ins.2011.10.011.

Manjón, J.V., Tohka, J., Robles, M., 2010. Improved estimates of partial volume coefficients from noisy brain MRI using spatial context. Neuroimage 53 (2), 480–490. doi:10.1016/j.neuroimage.2010.06.046.

Pantoni, L., The LADIS Study group, 2011. 2001–2011: A decade of the LADIS (leukoaraiosis and DISability) study: what have we learned about white matter changes and small-vessel disease. Cerebrovascular Dis. 32, 577–588. doi:10.1159/000334498.

Prins, N.D., Scheltens, P., 2015. White matter hyperintensities, cognitive impairment and dementia: an update. Nat. Rev. Neurol..

Prins, N.D., van Straaten, E.C.W., van Dijk, E.J., Simoni, M., van Schijndel, R.a., Vrooman, H.a., Koudstaal, P.J., Scheltens, P., Breteler, M.M.B., Barkhof, F., 2004. Measuring progression of cerebral white matter lesions on MRI: visual rating and volumetrics. Neurology 62 (9), 1533–1539. doi:10.1212/01.WNL.0000123264.40498.B6.

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.

Rey, D., Subsol, G., Delingette, H., Ayache, N., 2002. Automatic detection and segmentation of evolving processes in 3D medical images: application to multiple sclerosis. Med. Image Anal. 6 (2), 163–179. doi:10.1016/S1361-8415(02)00056-7.

Schmahmann, J.D., Smith, E.E., Eichler, F.S., Filley, C.M., 2008. Cerebral white matter: neuroanatomy, clinical neurology, and neurobehavioral correlates. Ann. N. Y. Acad. Sci. 1142, 266–309. doi:10.1196/annals.1444.017. NIHMS150003.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Mühlau, M., 2012. An automated tool for detection of FLAIR-hyperintense white matter lesions in multiple sclerosis. Neuroimage 59, 3774–3783.

Schmidt, R., Ropele, S., Enzinger, C., Petrovic, K., Smith, S., Schmidt, H., Matthews, P.M., Fazekas, F., 2005. White matter lesion progression, brain atrophy, and cognitive decline: the austrian stroke prevention study. Ann. Neurol. 58 (4). doi:10.1002/ana.20630. 610–6

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. Neuroimage 49 (2), 1524–1535.

Silbert, L.C., Nelson, C., Howieson, D.B., Moore, M.M., Kaye, J.A., 2008. Impact of white matter hyperintensity volume progression on rate of cognitive and motor decline. Neurology 71 (2), 108–113. doi:10.1212/01.wnl.0000316799.86917.37.

Styner, M.A., Lee, J., Chin, B., Chin, M.S., Commowick, O., Tran, H.-H., Markovic–Plese, S., Jewells, V., Warfield, S.K., 2008. 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. The MIDAS Journal - MS Lesion Segmentation.

Sudre, C., Cardoso, M.J., Bouvy, W., Biessels, G., Barnes, J., Ourselin, S., 2015. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation.. IEEE Trans. Med. Imaging 34 (10), 2079–2102. doi:10.1109/TMI.2015.2419072.

Targosz-Gajniak, M., Siuda, J., Ochudo, S., Opala, G., 2009. Cerebral white matter lesions in patients with dementia – from MCI to severe Alzheimer's disease. J. Neurol. Sci. 283 (1–2), 79–82. doi:10.1016/j.jns.2009.02.314.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20 (8), 677–688. doi:10.1109/42.938237.

Vuorinen, M., Solomon, A., Rovio, S., Nieminen, L., Kåreholt, I., Tuomilehto, J., Soininen, H., Kivipelto, M., 2011. Changes in vascular risk factors from midlife to late life and white matter lesions: a 20-year follow-up study. Dement Geriatr. Cogn. Disord. 31 (2). doi:10.1159/000323810. 119–25.

Wack, D.S., Dwyer, M.G., Bergsland, N., Di Perri, C., Ranza, L., Hussein, S., Ramasamy, D., Poloni, G., Zivadinov, R., 2012. Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates. BMC Med. Imaging 12 (17).

Wakefield, D.B., Moscufo, N., Guttmann, C.R., Kuchel, G.A., Kaplan, R.F., Pearlson, G., Wolfson, L., 2010. White matter hyperintensities predict functional decline in voiding, mobility, and cognition in older adults. J. Am. Geriatr. Soc. 58 (2), 275–281. doi:10.1111/j.1532-5415.2009.02699.x.

Wardlaw, J.M., Smith, C., Dichgans, M., 2013. Mechanisms underlying sporadic cerebral small vessel disease: insights from neuroimaging. Lancet Neurol. 12 (5). doi:10.1016/S1474-4422(13)70060-7.Mechanisms.