An open-source tool for longitudinal whole-brain and white matter lesion segmentation

Stefano Cerri [a,*], Douglas N. Greve [a,c], Andrew Hoopes [a], Henrik Lundell [b], Hartwig R. Siebner [b,d,e], Mark Mühlau [f], Koen Van Leemput [a,g]

[a] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA
[b] Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Amager and Hvidovre, Copenhagen, Denmark
[c] Department of Radiology, Harvard Medical School, USA
[d] Department of Neurology, Copenhagen University Hospital Bispebjerg and Frederiksberg, Copenhagen, Denmark
[e] Institute for Clinical Medicine, Faculty of Medical and Health Sciences, University of Copenhagen, Denmark
[f] Department of Neurology and TUM-Neuroimaging Center, School of Medicine, Technical University of Munich, Germany
[g] Department of Health Technology, Technical University of Denmark, Denmark

ARTICLE INFO

ABSTRACT

In this paper we describe and validate a longitudinal method for whole-brain segmentation of longitudinal MRI scans. It builds upon an existing whole-brain segmentation method that can handle multi-contrast data and robustly analyze images with white matter lesions. This method is here extended with subject-specific latent variables that encourage temporal consistency between its segmentation results, enabling it to better track subtle morphological changes in dozens of neuroanatomical structures and white matter lesions. We validate the proposed method on multiple datasets of control subjects and patients suffering from Alzheimer's disease and multiple sclerosis, and compare its results against those obtained with its original cross-sectional formulation and two benchmark longitudinal methods. The results indicate that the method attains a higher test–retest reliability, while being more sensitive to longitudinal disease effect differences between patient groups. An implementation is publicly available as part of the open-source neuroimaging package FreeSurfer.

1. Introduction

Longitudinal imaging studies, in which subjects are scanned repeatedly over time, have several advantages over cross-sectional studies. Accordingly, longitudinal neuroimaging studies have provided valuable insights into temporal changes in healthy brain development (Giedd et al., 1999; Evans et al., 2006; Group, 2012; Choe et al., 2013; Mills et al., 2021) and aging (Scahill et al., 2003; Tamnes et al., 2013), as well as from neurodegenerative diseases such as Alzheimer's disease (AD) (Fox et al., 1996; Laakso et al., 1998; Du et al., 2001; Halliday, 2017) or multiple sclerosis (MS) (Audoin et al., 2006; Fisher et al., 2008). In most instances, a longitudinal study design increases statistical power compared to a cross-sectional design. Furthermore, only longitudinal studies allow for a reliable evaluation of interventions such as treatment effects. Most importantly, only longitudinal measures allow for monitoring the individual patient. In neuroimaging, preprocessing tools are commonly designed for cross-sectional data so that their use in

longitudinal data may not fully exploit the advantages of the longitudinal study design with the risk of an overestimation of statistical power or the need of a higher number of subjects, respectively.

Over the last few decades, many dedicated neuroimage analysis tools have been developed to handle longitudinal data. These methods aim to exploit the expected temporal consistency in longitudinal scans to obtain more sensitive measures of longitudinal changes than is possible by analyzing each time point separately. One class of algorithms is designed to detect changes between two consecutive time points without explicitly *segmenting* each scan. These methods work by subtracting the two images to highlight locations of change (Hajnal et al., 1995; Freeborough and Fox, 1997; Lemieux et al., 1998; Battaglini et al., 2014), or, more generally, by tracking corresponding voxel locations over time using nonlinear registration, and analyzing the estimated spatial deformations (Thirion and Calmon, 1999; Rey et al., 2002; Avants et al., 2007; Holland and Dale, 2011; Elliott et al., 2019). Another class of methods explicitly *segments* each time point in longitudinal scans,

---
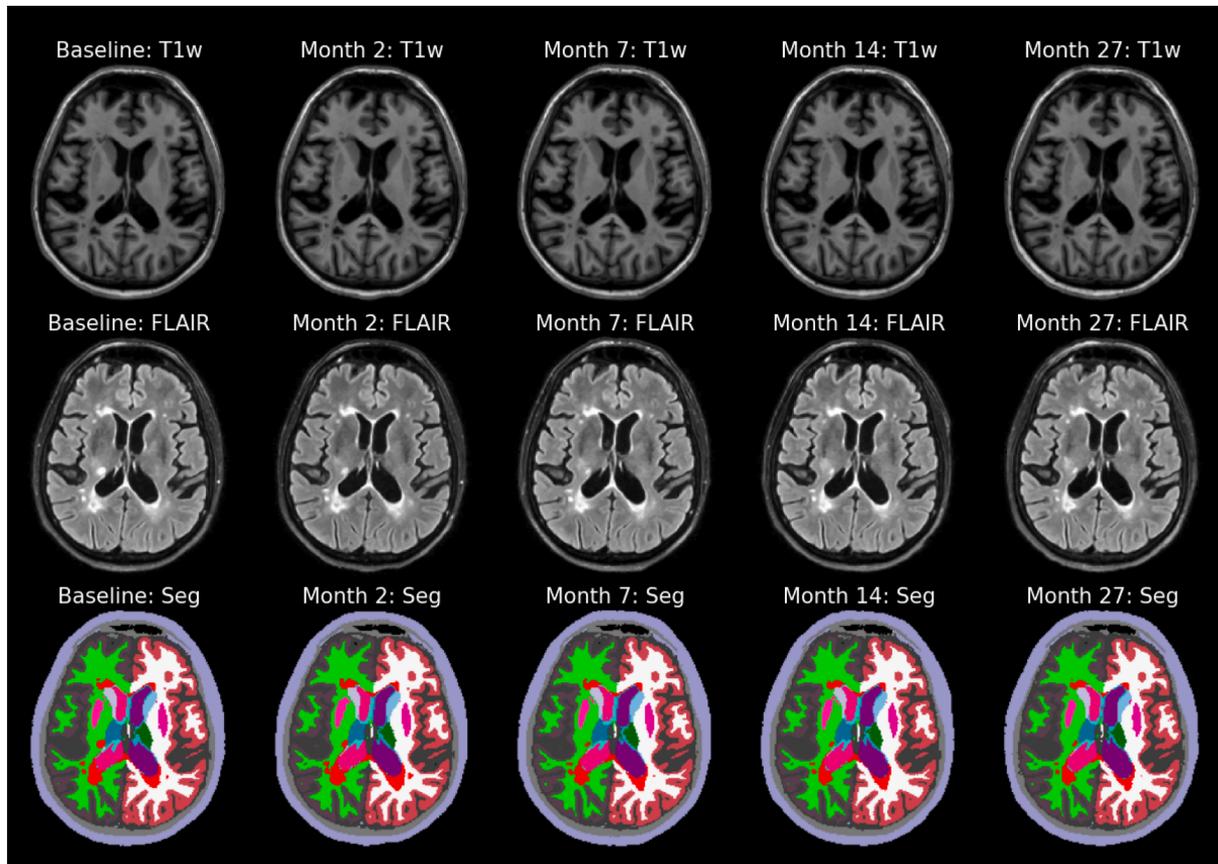
Fig. 1. Whole-brain and white matter lesion segmentations (labeled in red) produced by the proposed method from T1w and FLAIR longitudinal scans of an MS patient. T1w=T1-weighted, FLAIR=FLuid Attenuation Inversion Recovery, MS=Multiple Sclerosis.

enforcing temporal consistency either on the segmentations themselves (Metcalf et al., 1992; Solomon and Sood, 2004; Xue et al., 2005; Xue et al., 2006; Wolz et al., 2010; Dwyer et al., 2014; Wei et al., 2021) or on spatial probabilistic atlases that are used to compute them (Shi et al., 2010; Shi et al., 2010; Prastawa et al., 2012; Aubert-Broche et al., 2013; Iglesias et al., 2016; Tustison et al., 2019). In order to make the various time points comparable on a voxel-based level, these methods typically involve a temporal registration step, computed either prior to (Wolz et al., 2010; Aubert-Broche et al., 2013; Gao et al., 2014; Iglesias et al., 2016; Tustison et al., 2019; Schmidt et al., 2019; Wei et al., 2021) or simultaneously with (Xue et al., 2005; Xue et al., 2006; Shi et al., 2010; Shi et al., 2010; Li et al., 2010; Wang et al., 2011; Prastawa et al., 2012; Wang et al., 2013) the segmentations.

To date, most methods for analyzing longitudinal scans are designed to compute only very specific outcome variables, such as change in overall brain size (Hajnal et al., 1995; Freeborough and Fox, 1997; Smith et al., 2001; Smith et al., 2002) or global white/gray matter volume (Xue et al., 2005; Xue et al., 2006; Shi et al., 2010; Shi et al., 2010; Prastawa et al., 2012; Gao et al., 2014), cortical thickness (Nakamura et al., 2011; Wang et al., 2011; Wang et al., 2013; Tustison et al., 2019), white matter lesions (Gerig et al., 2000; Solomon and Sood, 2004; Elliott et al., 2013; Schmidt et al., 2019; Birenbaum and Greenspan, 2016; McKinley et al., 2020; Sepahvand et al., 2020; Denner et al., 2020) or individual brain structures such as the hippocampus (Wolz et al., 2010; Iglesias et al., 2016; Wei et al., 2021). To the best of our knowledge, the most comprehensive tool for longitudinal analysis of structural brain scans is currently the one distributed with FreeSurfer (Reuter et al., 2012; Fischl, 2012). This tool segments many neuroanatomical structures simultaneously (both volumetric "whole-brain" segmentations and parcellations of the cortical surface), and can readily handle data with more than two time points. However, it is specifically

designed for T1-weighted (T1w) scans only – as such it is less well suited for studying populations with white matter lesions and other pathologies that are better visualized using other MRI contrasts (such as T2w or FLAIR). Furthermore, a recent study suggests that, even in T1w images, it may be less sensitive to longitudinal changes than the method we describe here (Sederevičius et al., 2021).

The contribution of this paper is twofold. First, we make publicly available a new method for automatically segmenting dozens of neuroanatomical structures from longitudinal scans, using a model-based approach that can take multi-contrast data as input and that can also segment white matter lesions simultaneously. The method is fully adaptive to different MRI contrasts and scanners, and does not put any constraints on the number or the timing of longitudinal follow-up scans.

Second, we conduct an extensive validation of the proposed tool using over 4,500 brain scans acquired with different scanners, field strengths and acquisition protocols, involving both controls and patients suffering from MS and AD. We demonstrate experimentally that the method produces more reliable segmentations in scan-rescan settings than the longitudinal tool in FreeSurfer and than a cross-sectional version of the method, while also being more sensitive to differences in longitudinal changes between patient groups.

An example of longitudinal segmentations of an MS patient produced by the proposed method is shown in Fig. 1. A preliminary version of this work, with a limited validation, appeared earlier as a workshop paper (Cerri et al., 2020).

2. Existing cross-sectional method – SAMSEG

We build upon a previously validated cross-sectional method for whole-brain segmentation called Sequence Adaptive Multimodal SEG-mentation (SAMSEG) (Puonti et al., 2016). SAMSEG segments 41
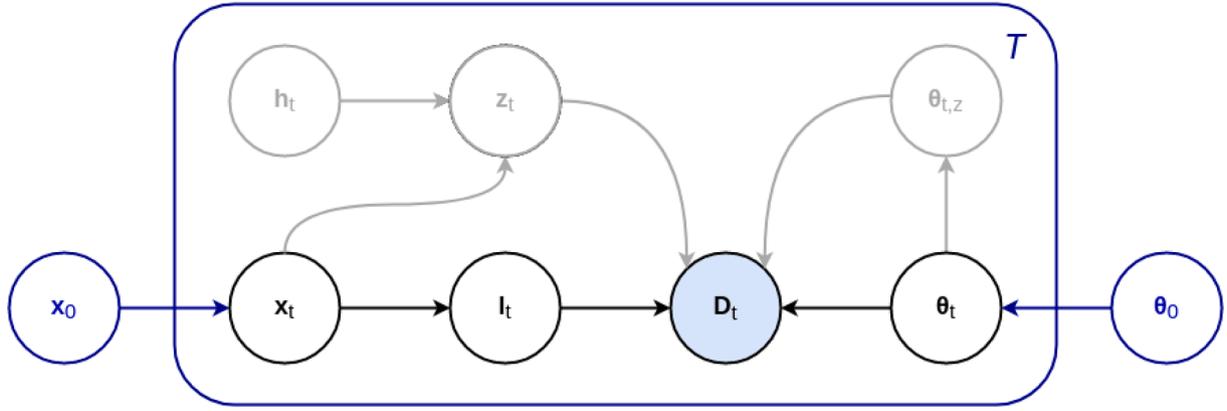
Fig. 2. Graphical representation of the proposed longitudinal generative model. For each time point t, in black the cross-sectional model of (Puonti et al., 2016), in which image intensities $\mathbf{D}_t$ are generated from likelihood parameters $\boldsymbol{\theta}_t$ and segmentation labels $\mathbf{l}_t$ – which in turn have been generated from an atlas-based segmentation prior with node positions $\mathbf{x}_t$. In blue, the proposed additional subject-specific latent variables that encourage temporal consistency between longitudinal scans in the segmentation prior (through $\mathbf{x}_0$) and in the likelihood function (through $\boldsymbol{\theta}_0$). Also shown, in gray, is the cross-sectional lesion extension of (Cerri et al., 2021). For each time point t, $\mathbf{z}_t$ is a binary white matter lesion segmentation, $\mathbf{h}_t$ are latent variables encoding lesion shape information, and $\boldsymbol{\theta}_{t,z}$ are lesion intensity parameters constraining lesion appearance. Shading indicates observed variables, while the plate indicates T repetition of the included variables.

anatomical structures from brain MRI, and it is fully adaptive to different MRI contrasts and scanners. We here briefly describe the method as we extend it for longitudinal scans in the remainder of the paper.

Let $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_I)$ be the image intensities of a multi contrast scan with I voxels, where $\mathbf{d}_i = (d_i^1, ..., d_i^N)^T$ is the vector containing the log-transformed image intensities of voxel i for all the available N contrasts. Furthermore, let $\mathbf{l} = (l_1, ..., l_I)^T$ be the corresponding segmentation labels, where $l_i \in \{1, ..., K\}$ denotes one of the K possible anatomical structures assigned to voxel i. In order to compute segmentation labels $\mathbf{l}$ from image intensities $\mathbf{D}$, we use a generative model illustrated in black in Fig. 2. It defines a forward model composed of two parts: a segmentation prior $p(\mathbf{l}|\mathbf{x})$, with parameters $\mathbf{x}$, that encodes spatial information of the labels $\mathbf{l}$, and a likelihood function $p(\mathbf{D}|\mathbf{l}, \boldsymbol{\theta})$, with parameters $\boldsymbol{\theta}$, that models the imaging process used to obtain the data $\mathbf{D}$. This forward model can be "inverted" to obtain automated segmentations, as detailed below.

2.1. Segmentation prior

We use a segmentation prior based on a deformable probabilistic atlas encoded as a tetrahedral mesh (Van Leemput, 2009). The mesh has node positions $\mathbf{x}$, governed by a deformation prior distribution defined as:

$$p(\mathbf{x}) \propto \exp\left[- \mathscr{K} \sum_{m=1}^{M} U_m\left(\mathbf{x}, \mathbf{x}_{ref}\right) \right], \tag{1}$$

where M is the number of tetrahedra in the mesh, $\mathscr{K} > 0$ controls the stiffness of the mesh, and $U_m(\mathbf{x}, \mathbf{x}_{ref})$ is a topology-preserving cost associated with deforming the mth tetrahedron from its shape in the atlas's reference position $\mathbf{x}_{ref}$ (Ashburner et al., 2000).

Given a deformed mesh with node positions $\mathbf{x}$, the probability $p(l_i = k|\mathbf{x})$ of observing label k at voxel i is obtained using baricentric interpolation. Assuming conditional independence of the labels between voxels finally yields

$$p(\mathbf{l}|\mathbf{x}) = \prod_{i=1}^{I} p(l_i|\mathbf{x}).$$

2.2. Likelihood function

We use a multivariate Gaussian intensity model for each of the K different structures, and model the bias field artifact as a linear combination of spatially smooth basis functions that is added to the local voxel intensities (Wells et al., 1996; Van Leemput et al., 1999). Letting $\boldsymbol{\theta}$ be the collection of the bias field parameters and intensity means and variances, the likelihood function is defined as

$$p(\mathbf{D}|\mathbf{l}, \boldsymbol{\theta}) = \prod_{i=1}^{I} p(\mathbf{d}_i|l_i, \boldsymbol{\theta}),$$

$$p(\mathbf{d}_i|l_i = k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{d}_i|\boldsymbol{\mu}_k + \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_k),$$

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_N^T \end{pmatrix}, \quad \mathbf{c}_n = \begin{pmatrix} c_{n,1} \\ \vdots \\ c_{n,P} \end{pmatrix}, \quad \boldsymbol{\phi}_i = \begin{pmatrix} \phi_1^i \\ \vdots \\ \phi_P^i \end{pmatrix},$$

where P denotes the number of bias field basis functions, ϕ_p^i is the basis function p evaluated at voxel i, and $\mathbf{c}_n$ collects the bias field coefficients for MRI contrast n. Furthermore, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the Gaussian mean and variance of structure k, respectively. A flat prior is used for the parameters of the likelihood, i.e., $p(\boldsymbol{\theta}) \propto 1$.

2.3. Segmentation

Given an MRI scan $\mathbf{D}$, a corresponding segmentation is obtained by first fitting the model to the data:

$$\widehat{\mathbf{x}}, \widehat{\boldsymbol{\theta}} = \underset{\mathbf{x}, \boldsymbol{\theta}}{\operatorname{argmax}} \, p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{D}). \tag{2}$$

The optimization problem in (2) is solved using a coordinate ascent scheme, in which $\mathbf{x}$ and then $\boldsymbol{\theta}$ are iteratively updated, each in turn. Once the model parameter estimates $\{\widehat{\mathbf{x}}, \widehat{\boldsymbol{\theta}}\}$ are available, the corresponding maximum a posteriori (MAP) segmentation is obtained as

$$\widehat{\mathbf{l}} = \underset{\mathbf{l}}{\operatorname{argmax}} \, p(\mathbf{l}|\mathbf{D}, \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{x}}).$$

Since $p(\mathbf{l}|\mathbf{D}, \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{x}}) \propto p(\mathbf{D}|\mathbf{l}, \widehat{\boldsymbol{\theta}})p(\mathbf{l}|\widehat{\mathbf{x}})$ and therefore factorizes over i, the optimal segmentation label can be computed for each voxel independently:

$$\widehat{l}_i = \underset{k}{\operatorname{argmax}} \frac{\mathcal{N}(\mathbf{d}_i|\widehat{\boldsymbol{\mu}}_k + \widehat{\mathbf{C}}\boldsymbol{\phi}_i, \widehat{\boldsymbol{\Sigma}}_k)p(l_i = k|\widehat{\mathbf{x}})}{\sum_{k'=1}^{K} \mathcal{N}(\mathbf{d}_i|\widehat{\boldsymbol{\mu}}_{k'} + \widehat{\mathbf{C}}\boldsymbol{\phi}_i, \widehat{\boldsymbol{\Sigma}}_{k'})p(l_i = k'|\widehat{\mathbf{x}})}. \tag{3}$$

More details can be found in (Puonti et al., 2016).

3. Longitudinal method – SAMSEG-Long

We now describe how we extend SAMSEG for longitudinal scans. In the remainder of the paper, we call the proposed longitudinal method *SAMSEG-Long*.

In a longitudinal scenario, we aim to compute automatic segmentations $\{\mathbf{l}_t\}_{t=1}^T$ from T consecutive scans with image intensities $\{\mathbf{D}_t\}_{t=1}^T$. In contrast to the cross-sectional setting where each image is treated independently, here we can exploit the fact that all images belong to the same subject to produce more consistent (and potentially more accurate) segmentations. Towards this end, we introduce subject-specific latent variables $\mathbf{x}_0$ and $\boldsymbol{\theta}_0$ in the segmentation prior and likelihood function of SAMSEG, respectively. The purpose of these additional components in the model – illustrated in blue in Fig. 2 – is to impose a statistical dependency between the time points, encouraging the segmentations to remain similar to one another.

In the following, we describe how the new subject-specific latent variables are defined in the segmentation prior and likelihood function, and how we obtain the corresponding segmentations accordingly. We will use the notation $\mathbf{x}_t$ and $\boldsymbol{\theta}_t$ to indicate the parameters of the prior and likelihood function at time t, respectively.

3.1. Segmentation prior

In order to obtain temporal consistency in the segmentation prior, we use the concept of a "subject-specific atlas" (Iglesias et al., 2016): a deformation of the cross-sectional atlas to represent the average subject-specific anatomy across all time points. In particular, we use

$$p\left(\{\mathbf{x}_t\}_{t=1}^T \middle| \mathbf{x}_0\right) = \prod_{t=1}^T p(\mathbf{x}_t|\mathbf{x}_0)$$

with

$$p(\mathbf{x}_t|\mathbf{x}_0) \propto \exp\left[-\mathscr{K}\sum_{m=1}^M U_m(\mathbf{x}_t,\ \mathbf{x}_0)\right],$$

where $\mathbf{x}_0$ are latent atlas node positions encoding subject-specific brain shape, with prior

$$p(\mathbf{x}_0) \propto \exp\left[-\mathscr{K}_0\sum_{m=1}^M U_m(\mathbf{x}_0,\ \mathbf{x}_{ref})\right].$$

Here the mesh stiffness $\mathscr{K}_0$ is a hyperparameter of the model, the value of which we determine empirically using cross-validation (cf. Section 4.1).

Note that this formulation of the longitudinal segmentation prior is very flexible, as it does not impose specific temporal trajectories (e.g., monotonic growth) on the anatomy of the subject. Furthermore, by using a very large value for its hyperparameter $\mathscr{K}_0$, $\mathbf{x}_0$ can be forced to remain close to $\mathbf{x}_{ref}$, so that cross-sectional segmentation prior of (1) for each individual time point is retained as a special case.

3.2. Likelihood function

In a similar vein, we also introduce subject-specific latent variables to encourage temporal consistency in the Gaussian intensity models. For each anatomical structure k, we condition its Gaussian parameters $\{\boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_{t,k}\}_{t=1}^T$ on latent variables $\{\boldsymbol{\mu}_{0,k}, \boldsymbol{\Sigma}_{0,k}\}$ using a normal-inverse-Wishart (NIW) distribution:

$$p\left(\{\boldsymbol{\theta}_t\}_{t=1}^T \middle| \boldsymbol{\theta}_0\right) = \prod_{t=1}^T p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_0)$$

with

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_0) \propto \prod_{k=1}^K \mathscr{N}\left(\boldsymbol{\mu}_{t,k}\middle|\boldsymbol{\mu}_{0,k}, P_{0,k}^{-1}\boldsymbol{\Sigma}_{t,k}\right)\mathrm{IW}\left(\boldsymbol{\Sigma}_{t,k}\middle|P_{0,k}\boldsymbol{\Sigma}_{0,k}, P_{0,k} - N - 2\right),$$

where $\boldsymbol{\theta}_0 = \{\boldsymbol{\mu}_{0,k}, \boldsymbol{\Sigma}_{0,k}\}_{k=1}^K$ collects the latent variables of all structures, and is assumed to have a flat prior: $p(\boldsymbol{\theta}_0) \propto 1$. The effect of this longitudinal model is to encourage the means and variances of each structure to remain similar to some "prototype" $\boldsymbol{\mu}_{0,k}$ and $\boldsymbol{\Sigma}_{0,k}$, respectively, without having to specify a priori what values these prototypes should take. The strength of this effect is governed by a hyperparameter $P_{0,k} \geqslant 0$ for each structure, which we determine empirically using cross-validation (cf. Section 4.1).

Note that no temporal regularization is added to the parameters of the bias field model, since the bias field will typically vary between MRI sessions. (Differences in global intensity scaling between time points are automatically included in the bias field model as well.) Furthermore, by choosing hyperparameters $P_{0,k} = 0$ the temporal regularization of the Gaussian parameters can be switched off, in which case the proposed likelihood function devolves into that of the cross-sectional SAMSEG method for each time point separately (Section 2.2).

3.3. Segmentation

As in the cross-sectional case, segmentations are obtained by first fitting the model to the data:

$$\widehat{\boldsymbol{\theta}}_0, \widehat{\mathbf{x}}_0, \{\widehat{\mathbf{x}}_t, \widehat{\boldsymbol{\theta}}_t\}_{t=1}^T = \operatorname*{argmax}_{\boldsymbol{\theta}_0, \mathbf{x}_0, \{\mathbf{x}_t, \boldsymbol{\theta}_t\}_{t=1}^T} p\left(\boldsymbol{\theta}_0, \mathbf{x}_0, \{\mathbf{x}_t, \boldsymbol{\theta}_t\}_{t=1}^T \middle| \{\mathbf{D}_t\}_{t=1}^T\right). \tag{4}$$

We optimize (4) with a coordinate ascent scheme, where we iteratively update each variable one at the time. Because $p(\mathbf{x}_t|\mathbf{x}_0)$ is of the same form as the cross-sectional segmentation prior, and the NIW distribution used in $p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_0)$ is the conjugate prior for the mean and variance of a Gaussian distribution, estimating $\mathbf{x}_t$ and $\boldsymbol{\theta}_t$ from $\mathbf{D}_t$ for given values of $\mathbf{x}_0$ and $\boldsymbol{\theta}_0$ simply involves performing an optimization of the form of (2) for each time point t separately. Conversely, for given values $\{\mathbf{x}_t, \boldsymbol{\theta}_t\}_{t=1}^T$ the update for $\boldsymbol{\theta}_0$ is given in closed form:

$$\boldsymbol{\mu}_{0,k} \leftarrow \left(\sum_{t=1}^T \boldsymbol{\Sigma}_{t,k}^{-1}\right)^{-1} \sum_{t=1}^T \boldsymbol{\Sigma}_{t,k}^{-1}\boldsymbol{\mu}_{t,k},$$

$$\boldsymbol{\Sigma}_{0,k}^{-1} \leftarrow \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{\Sigma}_{t,k}^{-1}\right)\frac{P_{0,k}}{P_{0,k} - N - 2},$$

whereas updating $\mathbf{x}_0$ involves the optimization

$$\operatorname*{argmin}_{\mathbf{x}_0} \sum_{m=1}^M \left[\mathscr{K}_0 U_m(\mathbf{x}_0,\ \mathbf{x}_{ref}) + \mathscr{K}\sum_{t=1}^T U_m(\mathbf{x}_t,\ \mathbf{x}_0)\right],$$

which we solve numerically using a limited-memory BFGS algorithm.

Once all parameters are estimated, we obtain segmentations as described in the cross-sectional setting, i.e., by using (3) for each time point separately.

3.4. Extension for white matter lesion segmentation

We have previously extended the cross-sectional SAMSEG method using an augmented model that allows it to robustly segment lesions in the white matter (Cerri et al., 2021). In this extended version, lesions are modeled with an extra Gaussian in the likelihood function, and with lesion-specific location and shape constraints in the segmentation prior. These additional lesion components are illustrated in gray in the graphical model of Fig. 2; we refer the reader to (Cerri et al., 2021) for an in-depth description.

The SAMSEG version with lesion segmentation ability can easily be integrated into SAMSEG-Long, as illustrated in Fig. 2. Due to the highly
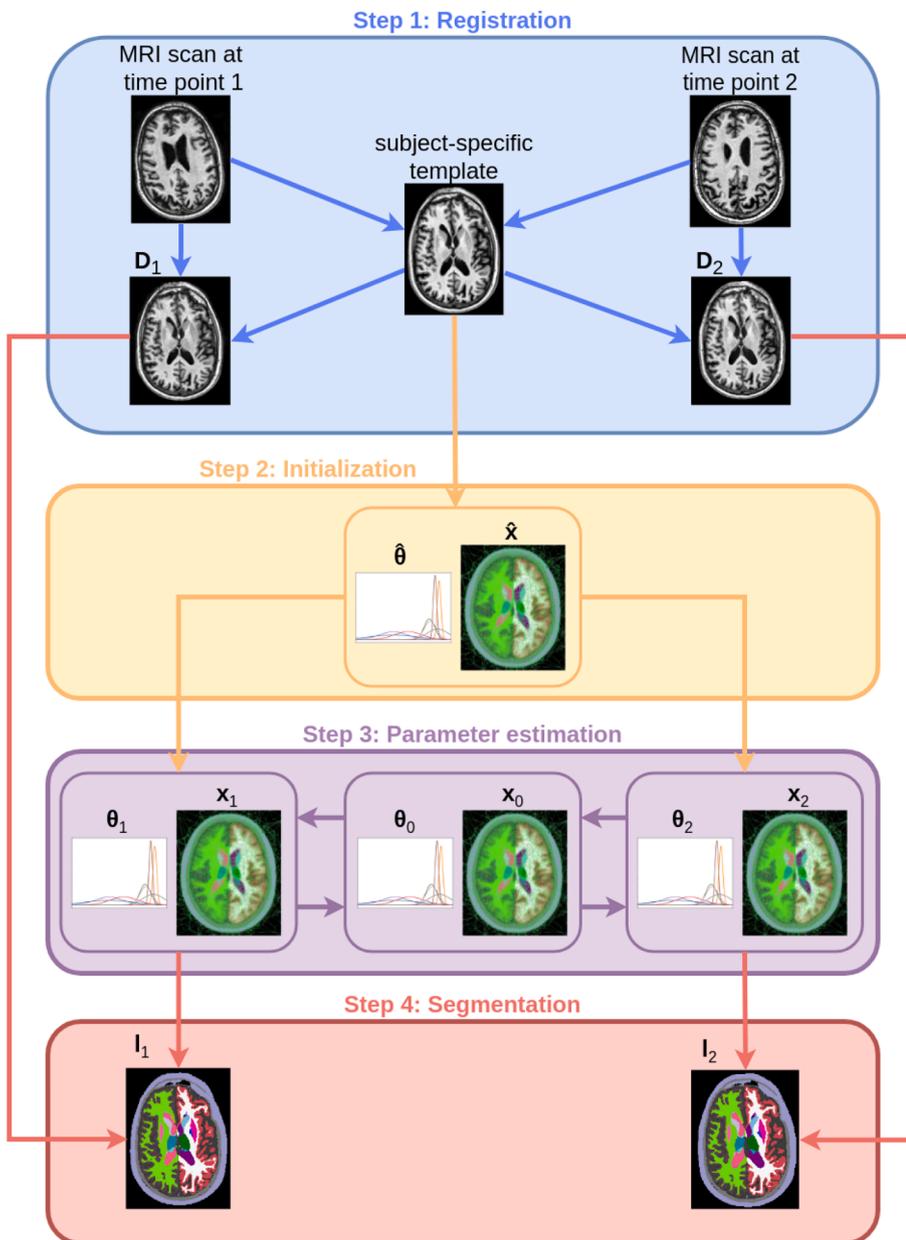
Fig. 3. Schematic illustration of the entire longitudinal segmentation process. Although the method is more general, here only the case of single-contrast scans and $T = 2$ time points is shown to avoid cluttering. In Step 1, an unbiased subject-specific template is created with an inverse consistent registration method (Reuter et al., 2012), and each input scan is subsequently registered to it (Section 4). The cross-sectional method (Section 2) is then applied to the template in Step 2, and the estimated parameters $\hat{\mathbf{x}}$ and $\hat{\theta}$ are used to initialize the corresponding parameters $\mathbf{x}_t$ and θ_t at each time point t. In Step 3, model parameters are estimated in an iterative process (Section 3.3); this involves alternating updates for subject-specific latent variables $\mathbf{x}_0$ and θ_0, with updates for time point parameters $\{\mathbf{x}_t, \theta_t\}_{t=1}^{T}$. Once model parameters estimates are available, in Step 4 each time point is segmented accordingly (Eq. (3)).

varying temporal behavior of white matter lesions, we do not explicitly constrain their shape and appearance over time, as this could potentially degrade the segmentation performance of the method. As a result, there are no direct dependencies between the lesion-specific components of the model and the latent variables $\mathbf{x}_0$ and θ_0 (note the absence of direct arrows between the two sets of variables in Fig. 2). Computing longitudinal whole-brain and white matter lesion segmentations can therefore follow the same procedure described before (cf. Section 3.3), with only a few modifications in how, for each time point t, the parameters $\mathbf{x}_t$ and θ_t are estimated, and individual segmentations $\mathbf{l}_t$ are computed (Cerri et al., 2021).

4. Implementation

In our current implementation, it is assumed that all time points have been registered and resampled to the same image grid prior to segmentation. To avoid introducing spurious biases by not treating all time points in exactly the same way (e.g., by resampling follow-up scans to a baseline scan) (Reuter and Fischl, 2011), for this purpose we use an unbiased within-subject template created with an inverse consistent registration method (Reuter et al., 2012). This template is a robust representation of the average subject anatomy over time, and we use it as an unbiased reference to register all time points to, resulting in resampled images that then form the input to our segmentation algorithm. In case of multi-contrast images, this procedure is performed for one specific contrast (T1w in the experiments used in this paper), and the remaining contrasts (FLAIR in the experiments) are subsequently registered and resampled to the first contrast for each time point individually.

To initialize the proposed algorithm, we first apply the cross-sectional method to the unbiased template, and use the estimated model parameters $\hat{\mathbf{x}}$ and $\hat{\theta}$ to initialize the corresponding parameters $\mathbf{x}_t$ and θ_t at each time point t. The model fitting procedure of (4), which interleaves updating the latent variables $\{\mathbf{x}_0, \theta_0\}$ with updating the parameters $\{\mathbf{x}_t, \theta_t\}_{t=1}^{T}$, is then run for five iterations, which we have found to be sufficient to reach convergence.

An illustration of the entire longitudinal segmentation process is provided in Fig. 3. Our implementation builds upon the C++ and Python

Table 1

Summary of the experiments and datasets used in the paper. CN=Cognitive Normal, CV=Converted, AD=Alzheimer's Disease, MS=Multiple Sclerosis, S-MS=Stable-MS, P-MS=Progressive-MS, # scans=total number of scans, tp=time points, avg-tp=average number of time points per subject, time-tp=average time in days between each time point, HT=Hyperparameter Tuning, TR=Test-Retest, IR-FSPGR=Inversion Recovery prepared - Fast SPoiled Gradient Recalled, MP-RAGE=Magnetization Prepared - RApid Gradient Echo, FLAIR=FLuid Attenuation Inversion Recovery. For more details about each individual dataset, see Appendix A.

Experiment	Dataset	Subjects	# scans	avg-tp (min, max)	time-tp (min, max)	Scanner	Sequence
Hyperparameter tuning	MIRIAD-TR-HT	CN=10, AD=30	80	2 (2, 2)	0 (0, 0)	GE Signa 1.5T	IR-FSPGR
	ADNI-HT	CN=37, AD=53	285	3.56 (2,5)	313 (107, 1121)	Multiple 3T scanners	MP-RAGE
							IR-FSPGR
Test–retest reliability	MIRIAD-TR	CN=13, AD=16	146	2 (2, 2)	0 (0, 0)	GE Signa 1.5T	IR-FSPGR
	OASIS-TR	CN=72, CV=14, AD=64	1845	3.55 (3, 4)	0 (0, 0)	Siemens Vision 1.5T	MP-RAGE
	Munich-TR	MS=2	34	5.67 (5, 6)	3 (2, 7)	Philips Achieva 3T Siemens Verio 3T GE Signa MR750 3T	IR-FSPGR MP-RAGE FLAIR
Detecting disease effects	ADNI	CN=66, AD=64	477	3.70 (2, 5)	298 (65, 903)	Multiple 1.5T and 3T scanners	IR-FSPGR MP-RAGE
	OASIS	CN=72, AD=64	336	2.47 (2, 5)	702 (182, 1510)	Siemens Vision 1.5T	MP-RAGE
	Munich	MS=200 (S-MS=100, P-MS=100)	1289	6.45 (2, 24)	353 (18, 3287)	Philips Achieva 3T	MP-RAGE FLAIR
Longitudinal lesion segmentation	ISBI	MS=14	61	4.36 (4, 6)	391 (299, 503)	Philips 3T	MP-RAGE FLAIR
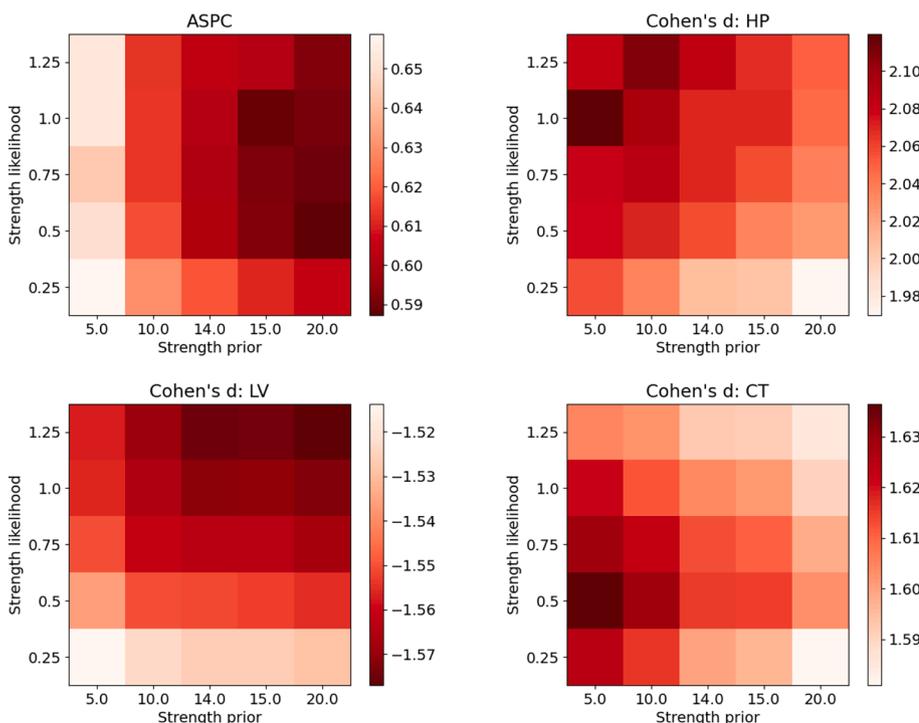
Fig. 4. Hyperparameter tuning by grid search over the hyperparameters of the proposed method. Top left: ASPC values across subjects and structures for the 80 test–retest T1w scans of the 40 subjects of the MIRIAD-TR-HT dataset (CN=10, AD=30). Top right and bottom row: Cohen's effect sizes computed from APCs estimates of the 80 subjects of the ADNI-HT dataset (CN=37, AD=53) for Hippocampus (HP), Lateral Ventricles (LV) and Cerebral Cortex (CT). APSC=Absolute Symmetrized Percent Change, CN=Cognitive Normal, AD=Alzheimer's Disease.

code of (Puonti et al., 2016 and Cerri et al., 2021), and is publicly available from FreeSurfer[1]. Segmenting one subject with 1 mm^3 isotropic resolution and image size of 256^3 takes approximately 10 min per time point for SAMSEG-Long, while 5 additional minutes per time point are needed when segmenting also white matter lesions (measured on an Intel 12-core i7-8700 K processor).

4.1. Hyperparameter tuning

SAMSEG-Long has hyperparameters $\mathscr{K}_0$ and $P_{0,k}$ that control,

respectively, the strength of the regularization in the segmentation prior and likelihood function. Good choices for the value of these hyperparameters will aim to minimize differences between scans acquired within a short interval of time, while simultaneously maximizing the ability to detect known atrophy trajectories in different patient groups.

We therefore tuned these hyperparameters by applying a grid search using 80 test–retest scans and 80 longitudinal scans of both cognitively normal (CN) and AD subjects. In particular, using T1w images from the MIRIAD-TR-HT (CN = 10, AD = 30) and the ADNI-HT (CN = 37, AD = 53) datasets summarized in Table 1 and detailed in Appendix A, we searched from the following values of the hyperparameters: $\mathscr{K}_0 = \{5\mathscr{K},\ 10\mathscr{K}, 14\mathscr{K}, 15\mathscr{K}, 20\mathscr{K}\}$ and $P_{0,k} = \{0.25N_k, 0.5N_k, 0.75N_k, N_k, 1.25N_k\}$, where N_k is the number of voxels assigned to class k in the cross-

---

[1] https://surfer.nmr.mgh.harvard.edu/fswiki/Samseg

sectional segmentation of the within-subject template.

The results are summarized in Fig. 4, which shows ASPC values in the test–retest scenario, as well as Cohen's d effect sizes of APC values between CN and AD patients for hippocampus, lateral ventricles and cerebral cortex – three structures known to be strongly affected in AD (Lombardi et al., 2020). The ASPC and APC metrics are defined in detail in Section 5.2, but in short assess volumetric changes between test and retest scans (ASPC), and the yearly rate of volume changes in longitudinal scans (APC), both expressed as a percentage.

The proposed method yielded consistent performance overall, with only minor differences between the various hyperparameter value combinations: ASPC and Cohen's d values varied within a 5.8% and 3.7% range compared to the average performance, respectively. Nevertheless, for the purpose of having fixed values for these hyperparameters, we used the combination $\mathscr{K}_0 = 20\mathscr{K}$ and $P_{0,k} = 0.5N_k, \forall k$ for all the experiments described below.

5. Experiments

In order to evaluate the performance of SAMSEG-Long, we conducted experiments on multiple datasets acquired with many different scanner platforms, field strengths, acquisition protocols, and image resolutions. These datasets contain images of cognitively normal subjects as well as AD and MS patients, and differ both in the number and timing of their longitudinal follow up scans, as well as in the number of MRI contrasts that are acquired. A summary of the datasets can be found in Table 1, with more detailed information for each individual dataset in Appendix A. Taken together, we believe these datasets are an excellent source to demonstrate the robustness and generalizability of the proposed longitudinal method, which does not need to be retrained or tuned on any of these datasets.

As a first experiment, we evaluated the method's test–retest reliability. Since test–retest scans are acquired within a minimal interval of time (usually within the same scan session or within a couple of weeks), no biological variations in the various structures are expected, and we therefore evaluated the ability of the method to produce consistent segmentations in these settings. Although this property is essential in a longitudinal segmentation method, an algorithm that produces the same segmentation for each given time point would, by definition, also have perfect test–retest reliability. Additional experiments are therefore needed to evaluate the ability of the method to also detect real longitudinal changes if they exist. Ideally, this would involve comparing the longitudinal segmentations computed by the proposed method with manually delineated longitudinal data. However, to the best of our knowledge, such ground truth data is not currently available. We therefore performed an *indirect* evaluation of the sensitivity of the method, by assessing its ability to detect known differences between the temporal trajectories in different patient groups (CN vs. AD, stable vs. progressive MS).

For all our experiments, we report the performance of the "vanilla" SAMSEG-Long method as described in Section 3.3 – except for MS patients, for which we use the method with its white matter lesion segmentation extension (Section 3.4).

5.1. Benchmark methods

In order to benchmark the longitudinal whole-brain segmentation performance of SAMSEG-Long, we compared it against that of SAMSEG (which is cross-sectional), and the longitudinal stream of FreeSurfer 7.2 (Reuter et al., 2012), called Aseg-Long in the remainder of the paper. Aseg-Long is the only publicly available and extensively validated longitudinal method that segments the same neuroanatomical structures as our method, representing a natural benchmark for evaluating its whole-brain segmentation performance. However, other tools exist that have reported better performance for estimating longitudinal volume changes in specific structures, such as the hippocampus, lateral ventricles or gray

matter (Nakamura et al., 2014; Guizard et al., 2015). Note that Aseg-Long is unable to process multi-contrast scans, hence no comparison was performed on such data.

For evaluating the lesion segmentation component of SAMSEG-Long, we compared its performance against that of both SAMSEG and the longitudinal white matter lesion segmentation method of (Schmidt et al., 2019), called LST-Long in the remainder of the paper. LST-Long is one the few methods that have a publicly available implementation. It segments lesions from T1-weighted and FLAIR MRI scans with multiple time points, and does not require retraining when tested on unseen data. Unlike our method, however, it does not provide further segmentations of the various neuroanatomical structures beyond white matter lesions.

5.2. Metrics and measures

Although the proposed method segments more structures, we concentrated on the following 25 main neuroanatomical regions: left and right cerebral white matter (WM), cerebellum white matter (CWM), cerebellum cortex (CCT), cerebral cortex (CT), lateral ventricle (LV), hippocampus (HP), thalamus (TH), putamen (PU), pallidum (PA), caudate (CA), amygdala (AM), nucleus accumbens (AC) and brain stem (BS). To avoid cluttering, we merged the results between right and left structures. For experiments that include MS patients, white matter lesion (LES) results are also reported.

For evaluating test–retest reliability, we computed the Absolute Symmetrized Percent Change (ASPC) for each structure, defined as

$$\text{ASPC} = \frac{100 \cdot |v_2 - v_1|}{(v_1 + v_2)/2},$$

where v_1 and v_2 represent the volume of the structure in scan 1 and scan 2, respectively. To check for under- and over-segmentation trends, we also computed Symmetrized Percent Change (SPC), defined in the same way as ASPC but without the absolute value.

For assessing a method's ability to detect disease effects, we computed the Annualized Percentage Change (APC): We fitted a line to the subject's volumetric measurements, plotted as a function of time from the baseline, and computed the APC as the ratio of its slope to its intercept (evaluated at the time of the first scan). A negative APC value thus corresponds to a yearly shrinkage, in percentage, of the structure of interest, while a positive APC value indicates a yearly growth. Effect sizes between APCs of two patient groups were then computed using Cohen's d (Cohen, 2013). We also report the minimum number of subjects needed to detect a statistically significant difference in atrophy rates between two patient groups, by first fitting a generalized linear model with APCs and corresponding patient groups. We then performed a power analysis using the computed model coefficients and noise variance, with 80% power and 0.05 significance level (Cohen, 2013).

For evaluating sensitivity to lesion activity in MS patients, we additionally assessed the apparent speed of lesion volume growth and shrinkage between segmentations at consecutive time points. Specifically, we computed annualized lesion volume *increase* (LES_I) and *decrease* (LES_D) as

$$\text{LES\_I} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{Nz_{t,t+1}}{\Delta_{t,t+1}}$$

and

$$\text{LES\_D} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{Nz_{t+1,t}}{\Delta_{t,t+1}},$$

respectively. Here $Nz_{t,t+1}$ counts the number of voxels that were labeled as lesion at time point $t+1$ but not at time point t, while $\Delta_{t,t+1}$ is the time passed between scan t and scan $t+1$. Intuitively, LES_I (LES_D) counts by how many voxels the total lesion volume has grown (shrunk) in a year, on average, *ignoring potential lesion areas where there was*
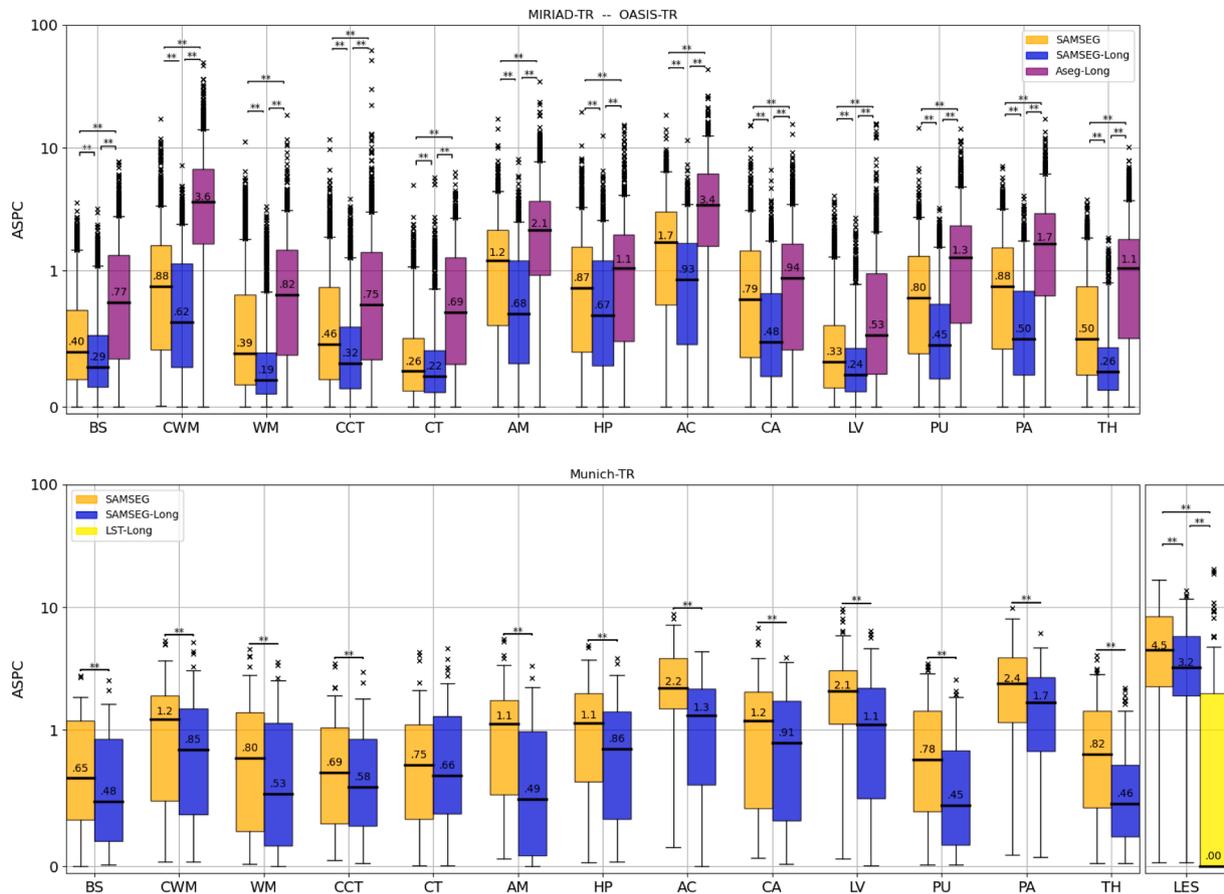
Fig. 5. ASPC values for the test–retest T1w scans of the combined MIRIAD-TR (# scans: 146, AD=13, CN=16) and OASIS-TR (# scans: 1845, AD=72, Converted=14, CN=64) datasets (top) and for the test–retest T1w-FLAIR scans of the Munich-TR (# scans: 34, MS=2) dataset (bottom). The yellow boxplot shows the result for LST-Long for white matter lesions. ASPC values were computed for different brain structures and white matter lesion (LES) for the proposed method and the benchmark methods. Within each boxplot, the median ASPC value is also reported. Statistically significant differences between two methods were computed with a Wilcoxon signed-rank test, and are indicated by asterisks ("**" indicates p-value < 0.01). ASPC=Absolute Symmetrized Percent Change, CN=Cognitive Normal, AD=Alzheimer's Disease, T1w=T1-weighted, FLAIR=FLuid Attenuation Inversion Recovery, MS=Multiple Sclerosis, BS=brain stem, CWM=cerebellum white matter, WM=cerebral white matter, CCT=cerebellum cortex, CT=cerebral cortex, AM=amygdala, HP=hippocampus, AC=nucleus accumbens, CA=caudate, LV=lateral ventricle, PU=putamen, PA=pallidum, TH=thalamus, LES=white matter lesion.

simultaneous shrinkage (growth). In MS, lesions as depicted by MRI not only appear and grow (Gaitán et al., 2011) but also shrink and disappear (Sethi et al., 2017; Battaglini et al., 2022), and these metrics have therefore been suggested as markers of disease activity (Pongratz et al., 2019).

To evaluate the lesion segmentation performance of the proposed method against ground truth delineations, we computed Dice coefficients as:

$$\mathrm{DICE}_{X,Y} = \frac{2|X \cap Y|}{|X| + |Y|},$$

where X and Y denote segmentation masks, and $|\cdot|$ counts the number of voxels in a mask.

6. Results

Throughout this section, we will refer to specific datasets with their names as defined in Table 1. As a mnemonic, datasets used to evaluate *test–retest* reliability have an affix "-TR" in their names. Whenever boxplots are used, the median is indicated by a horizontal line, plotted inside boxes that extend from the first to the third quartile values of the data. The range of the data is indicated by whiskers extending from the boxes, with outliers represented by "x" symbols.

6.1. Test–retest reliability

In order to evaluate if the proposed longitudinal method produces consistent segmentations over time, we assessed its performance on test–retest scans of three different datasets: 146 T1w test–retest scans of the 29 subjects of the MIRIAD-TR dataset (CN=13, AD=16), 1845 T1w test–retest scans of the 150 subjects of the OASIS-TR dataset (CN=72, Converted=14, AD=64) and 34 T1w and FLAIR test–retest scans of the 2 MS patients of the Munich-TR dataset. We thus computed automated segmentations for SAMSEG-Long and the benchmark methods, and assessed test–retest reliability performance in terms of ASPC values. When more than two test–retest scans were available for a subject, APSC values were computed for each possible combination of test–retest scan pairs. The results are shown in Fig. 5.

Since we observed similar results in the MIRIAD-TR and the OASIS-TR datasets, we here only report on their combined results. (We redirect the reader to Fig. B.1 for the results on the individual datasets.) On the combined MIRIAD-TR/OASIS-TR dataset, the median ASPC across all structures was best for SAMSEG-Long: 0.39, compared to 0.62 for SAMSEG and 1.12 for Aseg-Long. Similarly, on the Munich-TR dataset the median ASPC was 0.69 for SAMSEG-Long vs. 1.07 for SAMSEG (Aseg-Long cannot be run on this multi-contrasts dataset). The overall weaker performance on the Munich-TR scans can be explained by the fact that these were acquired within a 3-week interval, whereas the scans of the combined MIRIAD-TR/OASIS-TR dataset were acquired
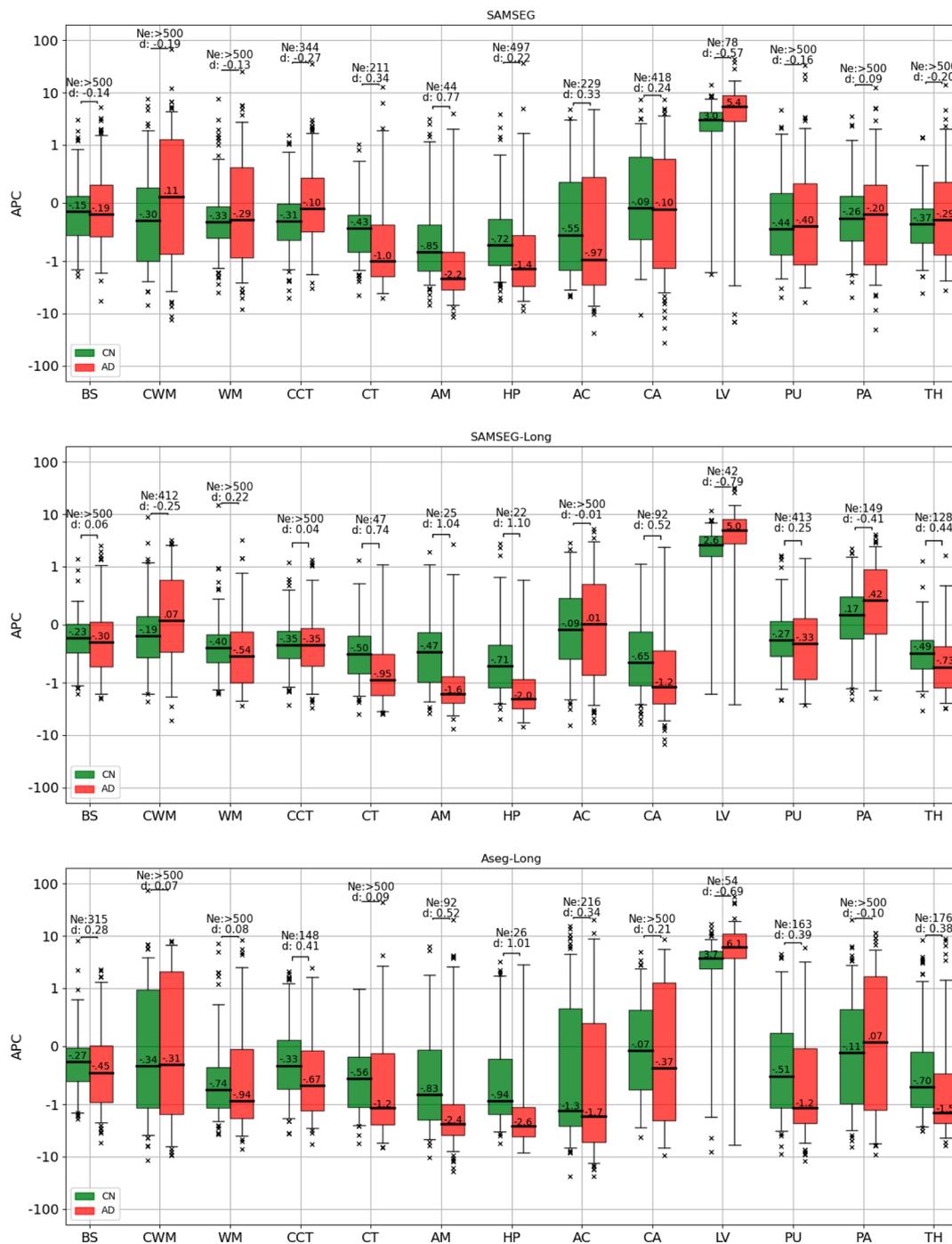
Fig. 6. APCs computed from the T1w scans of the 130 subjects of the ADNI dataset (CN=66, AD=64) and 136 subjects of the OASIS dataset (CN=72, AD=64) for SAMSEG, SAMSEG-Long, and Aseg-Long. Cohen's d effect size (d) and effective number of subjects (Ne) computed from a power analysis (80% power, 0.05 significance level) are reported above each pair of box plots. Within each boxplot, the median APC value is also indicated. APC=Annualized Percentage Change, CN=Cognitive Normal, AD=Alzheimer's Disease, T1w=T1-weighted, BS=brain stem, CWM=cerebellum white matter, WM=cerebral white matter, CCT=cerebellum cortex, CT=cerebral cortex, AM=amygdala, HP=hippocampus, AC=nucleus accumbens, CA=caudate, LV=lateral ventricle, PU=putamen, PA=pallidum, TH=thalamus.

within a single scan session without repositioning. Fig. 5 top shows a high number of outliers for the combined MIRIAD-TR/OASIS-TR dataset, with large ASPC values especially for Aseg-Long. This is mostly due to the high number of test–retest scans (approximately 2,000) of this combined dataset.

As for ASPC values of white matter lesions in the Munich-TR dataset (Fig. 5 bottom right), SAMSEG-Long outperformed SAMSEG (median ASPC: 4.5 vs. 3.2) even though the method does not explicitly regularize white matter lesions longitudinally (see Section 3.4). This may indicate

that more consistent model parameter estimates were obtained for SAMSEG-Long compared to SAMSEG as a result of enforcing temporal consistency on all the other structures. We also observe almost perfect test–retest reliability performance for LST-Long (median ASPC: 0.0), thereby outperforming both SAMSEG-LONG and SAMSEG by a large margin in this regard (but see below).

To check whether some of the methods are prone to under- or over-segmenting on test–retest scans, we also report the SPC values (i.e., without taking absolute values) on the same data. The results, shown in

Fig. 7. APCs of several structures computed from the T1w and FLAIR scans of the 200 patients of the Munich dataset (100 stable MS, and 100 progressive MS) for SAMSEG and SAMSEG-Long. For each comparison, Cohen's d effect size (d) and effective number of subjects (Ne) computed from a power analysis (80% power, 0.05 significance level) are shown above the boxplots. Within each boxplot, the median APC value is also reported. APC=Annualized Percentage Change, T1w=T1-weighted, FLAIR=FLuid Attenuation Inversion Recovery, MS=Multiple Sclerosis, BS=brain stem, CWM=cerebellum white matter, WM=cerebral white matter, CCT=cerebellum cortex, CT=cerebral cortex, AM=amygdala, HP=hippocampus, AC=nucleus accumbens, CA=caudate, LV=lateral ventricle, PU=putamen, PA=pallidum, TH=thalamus.
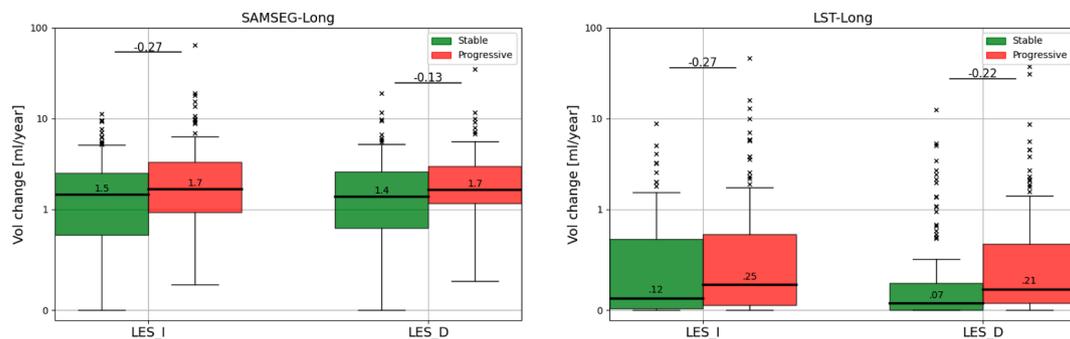
Fig. 8. Lesion volume increase (LES_I) and decrease (LES_D) computed from the T1w and FLAIR scans of the 200 patients of the Munich dataset (100 stable MS, and 100 progressive MS) for SAMSEG-Long and LST-Long. For each comparison, Cohen's d effect size is shown above the boxplots. Within each boxplot, the median value is also reported. T1w=T1-weighted, FLAIR=FLuid Attenuation Inversion Recovery, MS=Multiple Sclerosis.

Fig. B.2, did not indicate any particular trend in under- or over-segmenting specific structures for any of the methods (median SPC values for all the methods are close to 0), with SAMSEG-Long having the smallest SPC variances, followed by SAMSEG and Aseg-Long.

6.2. Detecting disease effects

Since we found similar findings across the two datasets, Fig. 6 shows the results for both datasets combined to ease readability. (We redirect the reader to Fig. B.3 and Fig. B.4 for individual dataset results.) All the methods were able to capture well-known differences in the atrophy trajectories of the hippocampus, amygdala and lateral ventricles between CN and AD patients (Lombardi et al., 2020). Effect sizes in these three structures differed between methods, with SAMSEG-Long having higher values for Cohen's d [0.79–1.10], followed by Aseg-Long [0.52–1.01] and SAMSEG [0.22–0.77]. The results of the power
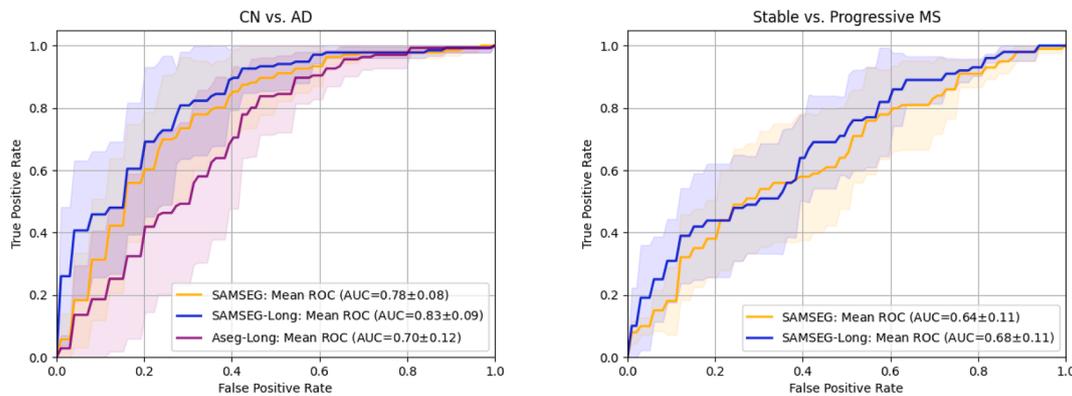
Fig. 9. ROC curves for an LDA classifier trained on APC values computed from the T1w scans of the ADNI (CN=66, AD=64) and OASIS (CN=72, AD=64) datasets (left) and from the T1w and FLAIR scans of the Munich dataset (MS stable=100, MS progressive=100) (right). LDA classifiers were trained and tested using a 5-fold cross-validation procedure. For each method, the area under the curve (AUC) is reported, the mean ROC curve is displayed as a solid line, and the shaded area represents ± 1 standard deviation error from the mean ROC curve. ROC=receiver operating characteristic curves, LDA=linear discriminant analysis, APC=Annualized Percentage Change, CN=Cognitive Normal, AD=Alzheimer's Disease, MS=Multiple Sclerosis, T1w=T1-weighted, FLAIR=FLuid Attenuation Inversion Recovery, MS=Multiple Sclerosis.
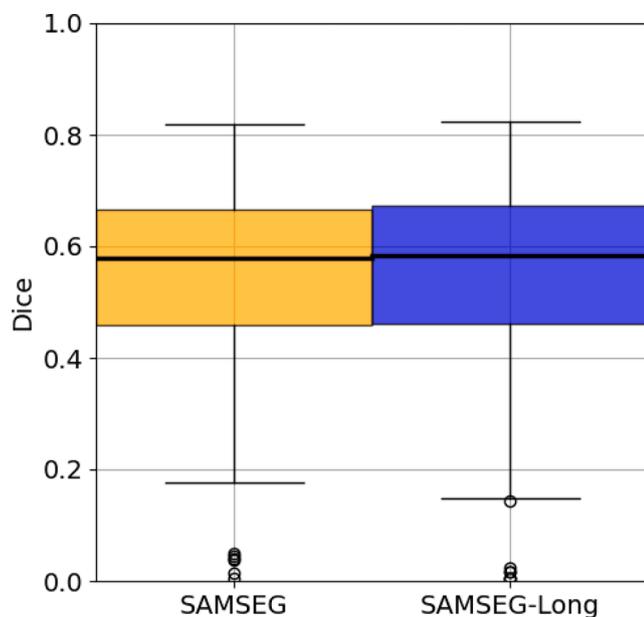
Fig. 10. Longitudinal lesion segmentation performance in terms of Dice overlaps with manual raters computed from T1w and FLAIR scans on the 14 MS subjects of the ISBI dataset for SAMSEG and SAMSEG-Long. Each automatic segmentation is compared against each of the two manual segmentations provided in the dataset. No statistically significant difference was detected between the two methods using a two-tailed paired t-test. T1w=T1-weighted, FLAIR=-FLuid Attenuation Inversion Recovery, MS=Multiple Sclerosis.

analysis closely mimick these findings: SAMSEG-Long requires fewer subjects to detect the differences in APC [22–42] compared to Aseg-Long [26–54] and SAMSEG [44–497]. Interestingly, only SAMSEG-Long showed a strong effect size for cerebral cortex (Cohen's d: 0.74), whose atrophy trajectories are known to differ between CN and AD patients (Edmonds et al., 2020), while both Aseg-Long and SAMSEG report lower effect sizes (Aseg-Long: 0.09, SAMSEG: 0.34).

Fig. 7 shows the same experiment for the 100 stable vs. the 100 progressive MS patients of the Munich dataset. Both SAMSEG-Long and SAMSEG yielded large differences in the atrophy trajectories of the cerebellum cortex, amygdala, hippocampus and thalamus, with SAMSEG-Long yielding nominally higher effect sizes and smaller sample sizes (higher power) compared to SAMSEG in these structures (Cohen's d: [0.37–0.53 vs. 0.33–0.50], sample sizes: [88–184 vs. 101–226]).

These results are in line with previous studies showing more marked atrophy trajectories in progressive MS patients than stable MS patients (Eshaghi et al., 2018; Cagol et al., 2022). We also report in Fig. 8 the yearly volume increase and decrease of lesions – as defined in Section 5.2, i.e., ignoring simultaneous shrinkage and growth, respectively – in stable and progressive MS patients, both for SAMSEG-Long and LST-Long. (Note that we cannot report such results for SAMSEG, as its segmentations are not longitudinally registered, therefore not allowing voxelwise lesion comparisons.) Similar to the findings in (Pongratz et al., 2019), where lesion volume increase and decrease were found to be comparable in size but larger in more active patients, both SAMSEG-Long and LST-Long detected more lesion changes (in both directions, i. e., lesion growth and shrinkage) in the progressive patients compared to the stable ones. Both methods yielded similar effect sizes between the two patient groups, but in absolute terms the lesion volume changes estimated by LST-Long were an order of magnitude smaller than the ones computed by SAMSEG-Long (0.07–0.25 ml/year vs. 1.4–1.7 ml/ year). Considering also the almost perfect test–retest performance of LST-Long in Section 6.1, it seems that the method may be over-regularizing over time.

As a final experiment to compare the various methods' ability to detect longitudinal disease effects, we assessed whether their APC values contain enough information to correctly classify individual subjects into their respective population groups. In contrast to previous experiments that focused on each structure independently, here we utilized all structures simultaneously: For each method, we trained and tested a linear discriminant analysis (LDA) classifier on APC values using a 5-fold cross-validation procedure. For each fold, 80% of the data was used for training and the remaining 20% for testing. Fig. 9 (left) shows receiver operating characteristic curves (ROC) obtained by training LDA classifiers on APC values computed from the ADNI and OASIS dataset for CN and AD patients. The LDA classifier trained on APC values computed by SAMSEG-Long achieved the highest area under the curve (0.83), followed by the classifiers trained on SAMSEG and Aseg-Long APC values (SAMSEG: 0.78, Aseg-Long: 0.70). Using the same cross-validation procedure, we also trained an LDA classifier on the APC values computed from the Munich dataset for stable and progressive MS patients, and reported ROC curves in Fig. 9 (right). In line with the previous experiment, the LDA classifier trained on SAMSEG-Long APC values obtained a higher area under the curve compared to the classifier trained on SAMSEG APC values (0.68 vs. 0.64).

6.3. Longitudinal lesion segmentation performance

In order to directly compare automatic longitudinal lesion segmentations against "ground truth" lesion annotations performed by human experts, we analyzed the 14 MS patients with longitudinal scans provided by the ISBI dataset. Since the heavy preprocessing applied to this data proved problematic for LST-Long, we were unable to get results with this method, and therefore only report performance for SAMSEG and SAMSEG-Long.

The ISBI challenge website[2] allows users to upload lesion segmentation masks, and ranks submissions according to an overall lesion segmentation performance score that takes into account Dice overlap, volume correlation, surface distance, and a few other metrics against manual annotations that remain hidden (see (Carass et al., 2017) for details). A score of 100 indicates perfect correspondence, while 90 is meant to correspond to human inter-rater performance (Carass et al., 2017; Styner et al., 2008). SAMSEG obtained a score of 88.31, while SAMSEG-Long similarly scored 88.61. The Dice coefficients between the manual and the corresponding automatic lesion segmentations – computed for each rater, subject and time point individually – are also provided by the website, and are summarized in Fig. 10. The median Dice score was around 0.58 for both SAMSEG and SAMSEG-Long, and no statistical significance was found between the two methods.

When interpreting these results, it is worth pointing out that, although the ISBI data itself is longitudinal, the *manual annotations* are not: As detailed in (Carass et al., 2017), the human raters were presented with the scan of each time point independently, resulting in poor longitudinal consistency in their manual annotations. Furthermore, both SAMSEG and SAMSEG-Long were used "out of the box" without any form of additional tuning. This should be taken into account when comparing their numerical scores against those obtained by methods that were specifically optimized for the exact scanner and imaging protocol of the challenge (typically by training on the matched training data that is also provided by the challenge).

7. Discussion and conclusion

In this paper we have proposed and evaluated a new method for segmenting dozens of neuroanatomical structures from longitudinal MRI scans. Temporal regularization is achieved by introducing a set of subject-specific latent variables in an existing cross-sectional segmentation method. An extension for segmenting white matter lesions is also available, allowing users to simultaneously track lesion evolution and morphological changes in various brain structures in e.g., patients suffering from MS. The proposed method does not make any assumptions on the scanner, the MRI protocol, or the number and timing of longitudinal follow-up scans, and is publicly available as part of the open-source neuroimaging package FreeSurfer.

Our experiments indicate that the proposed method has better test–retest reliability compared to benchmark methods, and that it is more sensitive to disease-related changes in AD and MS. In other words, our new tool generated results that were both more sensitive and more specific – suggesting that its use may bring several advantages, such as the need for fewer subjects in longitudinal studies, a better stratification of patients, and more precise evaluation of treatment efficacy.

The robustness and generalizability of the method across different scanner platforms, field strengths, acquisition protocols and image resolutions was demonstrated by its successful "out-of-the-box" application (i.e., without any form of retraining or retuning) on a diverse set of longitudinal datasets. These datasets included single- and multi-contrast longitudinal scans with a range of time gaps and total number of time points, from both healthy and diseased subjects, comprising over 4,500 MRIs in total. The cross-sectional methods the proposed technique builds upon have themselves previously been validated, by comparing their segmentations against those of manual raters on images acquired with different scanners and imaging protocols (Puonti et al., 2016; Cerri et al., 2021). Although this provides evidence of the generalizability of the method, a direct evaluation of its actual *longitudinal* segmentation performance has been hampered by the lack of manual longitudinal annotations that could serve as "ground truth". As is common in the longitudinal segmentation literature, we therefore resorted to an *indirect* validation and tested the ability of the method to detect disease-related temporal changes instead. This, however, has the limitation that higher sensitivity to group differences does not necessarily imply anatomically more correct segmentations.

Although we believe the proposed tool will be helpful for researchers and clinicians investigating temporal-morphological changes in the brain, the method still has several limitations. First, the method currently only produces volumetric whole-brain segmentations, as opposed to the longitudinal tool of (Reuter et al., 2012) that additionally also computes detailed longitudinal parcellations of the cortical surface. Second, although our tool can segment white matter lesions from longitudinal scans acquired with conventional MRI sequences, the signal changes that it detects in such images are nonspecific, with several different processes all resulting in similar MRI intensity profiles (Sethi et al., 2017; Pongratz et al., 2019). Disentangling the various underlying pathological changes due to e.g., demyelination, remyelination, inflammation or edema may ultimately become feasible with more advanced MRI techniques (Pirko and Johnson, 2008; Oh et al., 2019), but will likely require further development of the image analysis techniques described here. Third, the method does not currently exploit the time dimension explicitly to constrain (Metcalf et al., 1992; Welti et al., 2001; Solomon and Sood, 2004) or analyze lesion evolution (Thirion and Calmon, 1999; Gerig et al., 2000; Rey et al., 2002; Bosc et al., 2003; Elliott et al., 2013) in detail. Dedicated methods for detecting *new* or *growing* lesions by comparing two consecutive time points, in particular, have received ongoing attention (Commowick et al., 2021; Diaz-Hurtado et al., 2022): Many methods rely on image subtraction techniques (Sweeney et al., 2013; Battaglini et al., 2014; Ganiler et al., 2014; McKinley et al., 2020; Sepahvand et al., 2020; Krüger et al., 2020; Klistorner et al., 2021), while others use spatial deformation information (Elliott et al., 2019; Preziosa et al., 2022). Although this type of functionality is not directly provided by the proposed tool, its ability to tightly standardize longitudinal images – both in terms of removing global intensity scaling differences and bias field artifacts, and in terms of establishing accurate longitudinal nonlinear registrations across the various time points – may be leveraged to further develop such methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

---

[2] https://smart-stats-tools.org/lesion-challenge

Appendix A. Datasets details

We use disparate datasets for validating the proposed longitudinal method against benchmark methods, as well as for tuning the hyperparameters of the model. We here describe them in detail:

- **MIRIAD-TR-HT** (Malone et al., 2013): This dataset consists of test–retest scans of 10 healthy elderly people and 30 AD patients from the Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) project[3]. T1-weighted (T1w) scans with voxel size of $0.9375 \times 0.9375 \times 1.5$ mm were acquired on a GE Signa 1.5T scanner using an Inversion Recovery prepared - Fast SPoiled Gradient Recalled (IR-FSPGR) sequence. For each subject, 2 test–retest images were acquired without removing the patient from the scanner, for a total of 80 scans.
- **ADNI-HT**: This dataset consists of T1w longitudinal scans of 80 subjects from the ADNI project[4]. Different 3T scanners from multiple sites were used for scanning subjects. For each subject, T1w scans were acquired using an IR-FSPGR or MP-RAGE sequence at different image resolution ([min–max]: $[1.18–1.21] \times [0.92–1.31] \times [0.93–1.29]$ mm), with minor resolution variations between follow up scans ($\leqslant$0.01 mm in each axis). Subjects were scanned on average 3.56 times (minimum: 2 times; maximum: 5 times; 285 scans in total), with a mean interval between scans equal to 313 days (minimum: 107 days; maximum: 1121 days). The mean age at baseline was 75 years (minimum: 57; maximum: 90). Subjects were divided into two groups: cognitive normal (CN) (n=37) and AD (n=53).
- **MIRIAD-TR** (Malone et al., 2013): This dataset consists of test–retest scans of 13 healthy elderly people and 16 AD patients from the MIRIAD project – distinct from the subjects of the MIRIAD-TR-HT dataset. For each subject, 2 test–retest images were acquired at 2 or 3 different times without removing the patient from the scanner, for a total of 146 scans.
- **ADNI**: This dataset consists of T1w longitudinal scans of 130 subjects from the ADNI project – distinct from the subjects of the ADNI-HT dataset. Different scanners from multiple sites and multiple field strengths (1.5T and 3T) were used for scanning subjects. For each subject, T1w scans were acquired using an IR-FSPGR or MP-RAGE sequence at different image resolutions ([min–max]: $[1.17–1.21] \times [0.91–1.31] \times [0.92–1.31]$ mm), with minor resolution variations between follow-up scans ($\leqslant$0.01 mm in each axis). Subjects were scanned on average 3.70 times (minimum: 2 times; maximum: 5 times; 477 scans in total), with a mean interval between scans equal to 298 days (minimum: 65 days; maximum: 903 days). The mean age at baseline was 76 years (minimum: 57; maximum: 89). Subjects were divided into two groups: CN (n=66) and AD (n=64).
- **OASIS** (Marcus et al., 2010): This dataset consists of T1w longitudinal scans of 136 subjects from the Open Access Series of Imaging Studies (OASIS-2) project[5]. T1w scans with voxel size of $1 \times 1 \times 1.25$ mm were acquired on a 1.5T Siemens Vision scanner using a MP-RAGE sequence. Subjects were scanned on average 2.47 times (minimum: 2 times; maximum: 5 times; 336 scans in total), with a mean interval between scans equal to 702 days (minimum: 182 days; maximum: 1510 days). The mean age at baseline was 75 years (minimum: 60; maximum: 96). Scans were acquired on a Siemens Vision 1.5T scanner. Subjects were divided into two groups: CN (n = 72) and AD (n = 64).
- **OASIS-TR** (Marcus et al., 2010): This dataset consists of the same subjects of the OASIS dataset plus 14 additional subjects from the same project that were diagnosed as converted, for a total of 150 subjects. For each subject, 3 or 4 individual T1w MRI scans obtained in single scan sessions were acquired, for a total of 1845 scans.
- **Munich-TR** (Bieracher et al., 2016): This dataset consists of longitudinal T1w and FLuid Attenuation Inversion Recovery (FLAIR) scans of 2 MS subjects. For each subject, 6 repeated scans were acquired from 3 different 3T scanners (Philips Achieva; Siemens Verio; GE Signa MR750) within 3 weeks (mean interval between successive scans equal to 3 days (minimum: 2 days; maximum: 7 days). Voxel sizes for the T1w scans were $1 \times 1 \times 1$ mm for the GE scanner, $1 \times 1 \times 1$ mm for the Philips scanner and $1.1 \times 1.1 \times 1$ mm for the Siemens scanner. Voxel sizes for the FLAIR scans were $1 \times 1 \times 1$ mm for the GE scanner, $1 \times 1 \times 1.5$ mm for the Philips scanner, and $1 \times 1 \times 1$ mm for the Siemens scanner. T1w scans were acquired using a IR-FSPGR or MP-RAGE sequence. Two scans of "Subject-1" were excluded from the dataset due to scanning protocol violations (Bieracher et al., 2016), and they are not included in the experiments.
- **Munich**: This dataset consists of longitudinal T1w (MP-RAGE sequence) and FLAIR scans of 200 MS subjects acquired on a Philips Achieva 3T scanner. Subjects were scanned on average 6.45 times (minimum: 2 times; maximum: 24 times; 1289 scans in total), with a mean interval between

---

[3] https://www.nitrc.org/projects/miriad/

[4] http://adni.loni.usc.edu/

[5] https://www.oasis-brains.org/

scans equal to 353 days (minimum: 18 days; maximum: 3287 days) and at least 365 days between the first and last visit. Voxel sizes for the T1w scans were either $0.75 \times 0.75 \times 0.75$ or, after a scanner change, $1 \times 1 \times 1$ mm, while voxel sizes for the FLAIR scans were either $0.75 \times 0.75 \times 0.75$ or $0.9 \times 0.9 \times 1.5$ mm, respectively. Because some of the T1w scans have image resolutions that differ between time points of the same subject, we resampled all the T1w scans of this dataset to 1 mm isotropic resolution before using them in the experiments. The mean age at baseline was 40 years (minimum: 18; maximum: 74). At each subject's visit, an Expanded Disability Status Scale (EDSS) score was assessed (mean: 2.46; minimum: 0; maximum: 8.5). Subjects were divided into two groups: stable (n=100) and progressive (n=100) according to the following three strata criteria (Lorscheider et al., 2016): Patients were classified as progressive if (1) the baseline EDSS score is 0 and there is an increase of 1.5 points from the baseline EDSS score to the last time point EDSS score; (2) the baseline EDSS score is between 1 and 5.5 and there is an increase of 1 point from the baseline EDSS sscore to the last time point EDSS score; (3) the baseline EDSS score is above 5.5 and there is an increase of 0.5 points from the baseline EDSS score to the last time point EDSS score. Patients were classified as stable otherwise.

- **ISBI**: This dataset is the publicly available test set of the MS lesion segmentation challenge that was held at the 2015 International Symposium on Biomedical Imaging Carass et al. (2017). It consists of 14 longitudinal MS cases, scanned on average 4.36 times (minimum: 4 times, maximum: 6 times), separated by approximately one year (mean: 391 days, minimum: 299 days, maximum: 503 days). All images were acquired on the same Philips 3T scanner, using a 3D T1w, T2w, PDw and FLAIR sequence. Voxel sizes were $0.82 \times 0.82 \times 1.17$ mm for the T1w scans, and $0.82 \times 0.82 \times 2.22$ mm for T2w, PDw, and FLAIR. Images were first preprocessed (inhomogeneity correction, skull stripping, dura stripping, and again inhomogeneity correction – see (Carass et al., 2017) for details). Each T1w baseline image was then registered to a 1 mm MNI template and used as target image for registering successive time point images. Each subject's lesions were delineated by two different raters on the FLAIR scan, and, if necessary, corrected using the other contrasts. Note that the raters were presented with each scan independently, without aiming for longitudinal consistency in their segmentations. In our experiments we used the input combination T1w-FLAIR, which we have found to lead to good lesion segmentation performance in a previous study Cerri et al. (2021).

Appendix B. Additional results
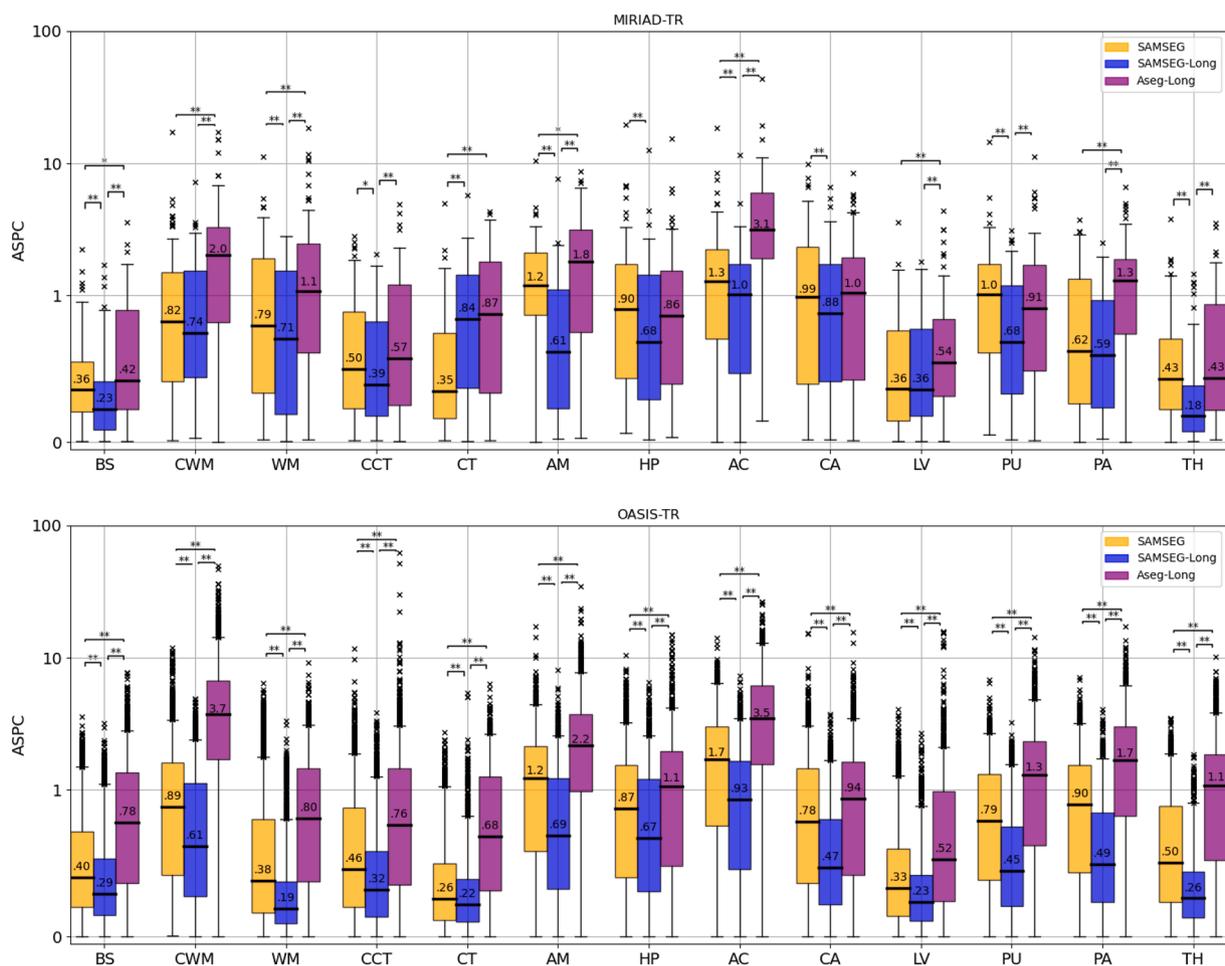
Fig. B.1. ASPCs for the test–retest T1w scans of the MIRIAD-TR (# scans: 146, AD=13, CN=16) and OASIS-TR (# scans: 1845, AD=72, Converted=14, CN=64) datasets for different brain structures for SAMSEG, SAMSEG-Long and Aseg-Long. Within each boxplot, the median ASPC value is also reported. Statistically significant differences between two methods, were computed with a Wilcoxon signed-rank test, and indicated by asterisks ("**" for p-value < 0.01 and "*" for p-value < 0.05). ASPC=Absolute Symmetrized Percent Change, T1w=T1-weighted, CN=Cognitive Normal, AD=Alzheimer's Disease, BS=brain stem, CWM=cerebellum white matter, WM=cerebral white matter, CCT=cerebellum cortex, CT=cerebral cortex, AM=amygdala, HP=hippocampus, AC=nucleus accumbens, CA=caudate, LV=lateral ventricle, PU=putamen, PA=pallidum, TH=thalamus.
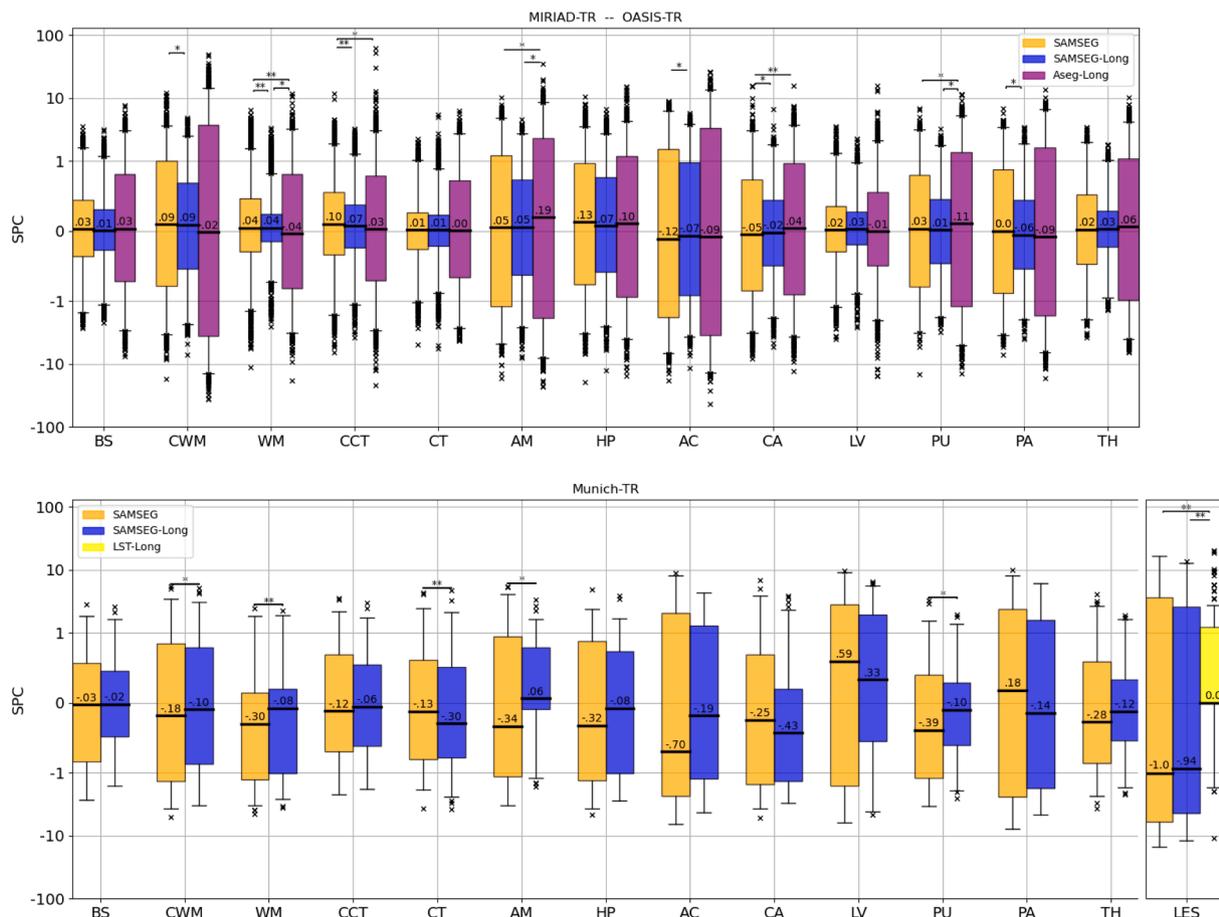
Fig. B.2. SPC values for the test–retest T1w scans of the combined MIRIAD-TR (# scans: 146, AD=13, CN=16) and OASIS-TR (# scans: 1845, AD=72, Converted=14, CN=64) datasets (top) and for the test–retest T1w-FLAIR scans of the Munich-TR (# scans: 34, MS=2) dataset (bottom). SPCs values were computed for different brain structures and white matter lesion (LES) for the proposed method and the benchmark methods. Within each boxplot, the median SPC value is also reported. Statistically significant differences between two methods were computed with a Wilcoxon signed-rank test, and are indicated by asterisks ("**" for p-value < 0.01, and "*" for p-value < 0.05). SPC=Symmetrized Percent Change, T1w=T1-weighted, FLAIR=FLuid Attenuation Inversion Recovery, AD=Alzheimer's Disease, CN=Cognitive Normal, MS=Multiple Sclerosis, BS=brain stem, CWM=cerebellum white matter, WM=cerebral white matter, CCT=cerebellum cortex, CT=cerebral cortex, AM=amygdala, HP=hippocampus, AC=nucleus accumbens, CA=caudate, LV=lateral ventricle, PU=putamen, PA=pallidum, TH=thalamus, LES=white matter lesion.
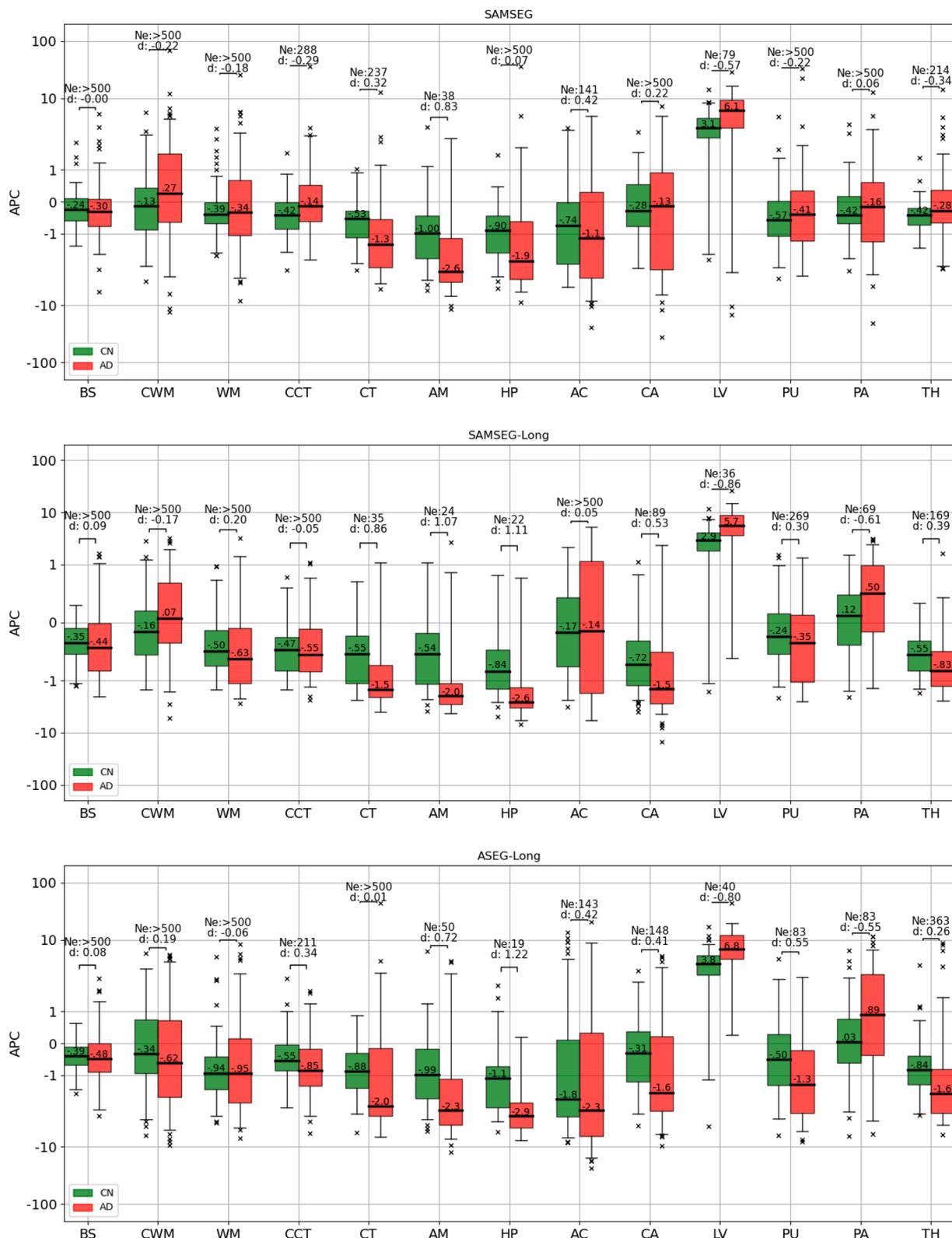
Fig. B.3. APCs computed from the T1W scans of the 130 subjects of the ADNI dataset (CN=66, AD=64) for SAMSEG, SAMSEG-Long, and Aseg-Long. Cohen's d effect size (d) and effective number of subjects (Ne) computed from a power analysis (80% power, 0.05 significance level) are reported above each pair of box plots. Within each boxplot, the median APC value is also reported. APC=Annualized Percentage Change, CN=Cognitive Normal, AD=Alzheimer's Disease, T1w=T1-weighted, BS=brain stem, CWM=cerebellum white matter, WM=cerebral white matter, CCT=cerebellum cortex, CT=cerebral cortex, AM=amygdala, HP=hippocampus, AC=nucleus accumbens, CA=caudate, LV=lateral ventricle, PU=putamen, PA=pallidum, TH=thalamus.
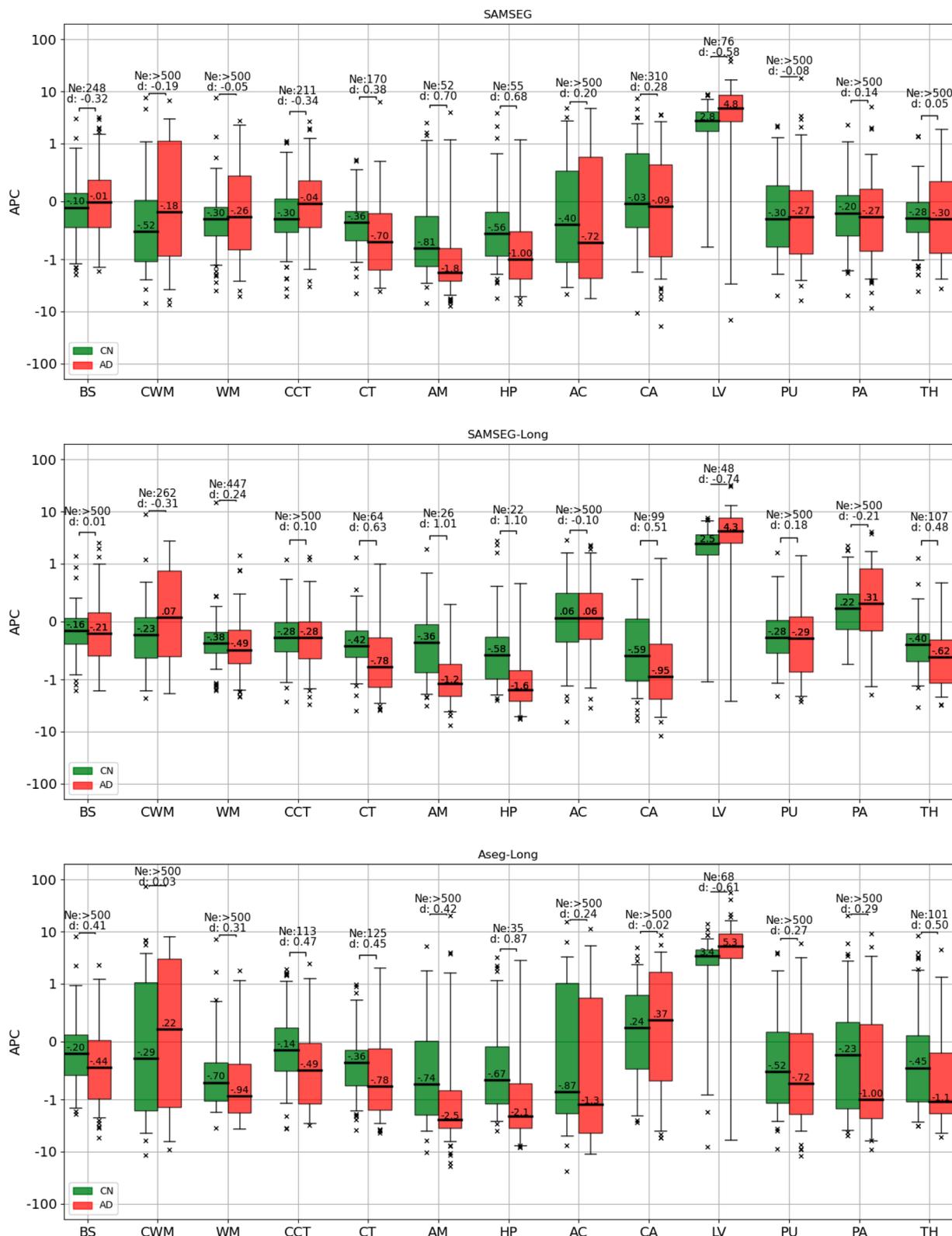
Fig. B.4. APCs computed from the T1W scans of the 136 subjects of the OASIS dataset (CN=72, AD=64) for SAMSEG, SAMSEG-Long, and Aseg-Long. Cohen's d effect size (d) and effective number of subjects (Ne) computed from a power analysis (80% power, 0.05 significance level) are reported above each pair of box plots. Within each boxplot, the median APC value is also reported. APC=Annualized Percentage Change, CN=Cognitive Normal, AD=Alzheimer's Disease, T1w=T1-weighted, BS=brain stem, CWM=cerebellum white matter, WM=cerebral white matter, CCT=cerebellum cortex, CT=cerebral cortex, AM=amygdala, HP=hippo-campus, AC=nucleus accumbens, CA=caudate, LV=lateral ventricle, PU=putamen, PA=pallidum, TH=thalamus.

References

Ashburner, J., Andersson, J.L., Friston, K.J., 2000. Image registration using a symmetric prior - In three dimensions. Hum. Brain Mapp. 9 (4), 212–225.

Aubert-Broche, B., Fonov, V.S., García-Lorenzo, D., Mouiha, A., Guizard, N., Coupé, P., Eskildsen, S.F., Collins, D.L., 2013. A new method for structural volume analysis of longitudinal brain MRI data and its application in studying the growth trajectories of anatomical brain structures in childhood. NeuroImage 82, 393–402.

Audoin, B., Davies, G.R., Finisku, L., Chard, D.T., Thompson, A.J., Miller, D.H., 2006. Localization of grey matter atrophy in early RRMS. J. Neurol. 253 (11), 1495–1501.

Avants, B., Anderson, C., Grossman, M., Gee, J.C., 2007. Spatiotemporal normalization for longitudinal analysis of gray matter atrophy in frontotemporal dementia. International Conference on Medical Image Computing and Computer-Assisted Intervention 4792, 303–310.

Battaglini, M., Rossi, F., Grove, R.A., Stromillo, M.L., Whitcher, B., Matthews, P.M., De Stefano, N., 2014. Automated identification of brain new lesions in multiple sclerosis using subtraction images. J. Magn. Reson. Imaging 39 (6), 1543–1549.

Battaglini, M., Vrenken, H., Tappa Brocci, R., Gentile, G., Luchetti, L., Versteeg, A., Freedman, M.S., Uitdehaag, B.M., Kappos, L., Comi, G., et al., 2022. Evolution from a first clinical demyelinating event to multiple sclerosis in the REFLEX trial: Regional susceptibility in the conversion to multiple sclerosis at disease onset and its amenability to subcutaneous interferon beta-1a. Eur. J. Neurol. 29 (7), 2024–2035.

Biberacher, V., Schmidt, P., Keshavan, A., Boucard, C.C., Righart, R., Sämann, P., Preibisch, C., Fröbel, D., Aly, L., Hemmer, B., Zimmer, C., Henry, R.G., Mühlau, M., 2016. Intra- and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. NeuroImage 142, 188–197.

Birenbaum, A., Greenspan, H., 2016. Longitudinal Multiple Sclerosis Lesion Segmentation Using Multi-view Convolutional Neural Networks. In: Deep Learning and Data Labeling for Medical Applications. Springer International Publishing, pp. 58–67.

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 (2), 643–656.

Cagol, A., Schaedelin, S., Barakovic, M., Benkert, P., Todea, R.-A., Rahmanzadeh, R., Galbusera, R., Lu, P.-J., Weigel, M., Melie-Garcia, L., et al., 2022. Association of Brain Atrophy With Disease Progression Independent of Relapse Activity in Patients With Relapsing Multiple Sclerosis, JAMA neurology.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Bazin, P.-L., Calabresi, P.A., Crainiceanu, C.M., Ellingsen, L.M., Reich, D. S., Prince, J.L., Pham, D.L., 2017. Longitudinal multiple sclerosis lesion segmentation data resource. Data in Brief 12, 346–350.

Cerri, S., Hoopes, A., Greve, D.N., Mühlau, M., Van Leemput, K., 2020. A Longitudinal Method for Simultaneous Whole-Brain and Lesion Segmentation in Multiple Sclerosis. Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology 12449, 119–128.

Cerri, S., Puonti, O., Meier, D.S., Wuerfel, J., Mühlau, M., Siebner, H.R., Van Leemput, K., 2021. A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. NeuroImage 225, 117471.

Choe, M.-S., Ortiz-Mantilla, S., Makris, N., Gregas, M., Bacic, J., Haehn, D., Kennedy, D., Pienaar, R., Caviness Jr, V.S., Benasich, A.A., et al., 2013. Regional infant brain development: an MRI-based morphometric analysis in 3 to 13 month olds. Cereb. Cortex 23 (9), 2100–2117.

Cohen, J., 2013. Statistical power analysis for the behavioral sciences. Routledge.

Commowick, O., Cervenansky, F., Cotton, F. and Dojat, M. (2021), MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure, in MICCAI 2021–24th International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 1–118.

Denner, S., Khakzar, A., Sajid, M., Saleh, M., Spiclin, Z., Kim, S.T., Navab, N., 2020. Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation. International MICCAI Brainlesion Workshop 111–121.

Diaz-Hurtado, M., Martínez-Heras, E., Solana, E., Casas-Roma, J., Llufriu, S., Kanber, B. and Prados, F. (2022), Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review, Neuroradiology pp. 1–15.

Du, A.T., Schuff, N., Amend, D., Laakso, M.P., Hsu, Y.Y., Jagust, W.J., Yaffe, K., Kramer, J.H., Reed, B., Norman, D., Chui, H.C., Weiner, M.W., 2001. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J. Neurol. Neurosurgery Psychiatry 71 (4), 441–447.

Dwyer, M.G., Bergsland, N., Zivadinov, R., 2014. Improved longitudinal gray and white matter atrophy assessment via application of a 4-dimensional hidden Markov random field model. Neuroimage 90, 207–217.

Edmonds, E.C., Weigand, A.J., Hatton, S.N., Marshall, A.J., Thomas, K.R., Ayala, D.A., Bondi, M.W., McDonald, C.R., Initiative, A.D.N., et al., 2020. Patterns of longitudinal cortical atrophy over 3 years in empirically derived MCI subtypes. Neurology 94 (24), e2532–e2544.

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.

Elliott, C., Wolinsky, J.S., Hauser, S.L., Kappos, L., Barkhof, F., Bernasconi, C., Wei, W., Belachew, S., Arnold, D.L., 2019. Slowly expanding/evolving lesions as a magnetic resonance imaging marker of chronic active multiple sclerosis lesions. Multiple Sclerosis J. 25 (14), 1915–1925.

Eshaghi, A., Marinescu, R.V., Young, A.L., Firth, N.C., Prados, F., Jorge Cardoso, M., Tur, C., De Angelis, F., Cawley, N., Brownlee, W.J., et al., 2018. Progression of regional grey matter atrophy in multiple sclerosis. Brain 141 (6), 1665–1677.

Evans, A.C., Group, B.D.C. et al. (2006), The NIH MRI study of normal brain development, Neuroimage 30(1), 184–202.

Fischl, B., 2012. FreeSurfer. NeuroImage 62 (2), 774–781.

Fisher, E., Lee, J.-C., Nakamura, K., Rudick, R.A., 2008. Gray matter atrophy in multiple sclerosis: a longitudinal study. Ann. Neurol.: Off. J. Am. Neurol. Assoc. Child Neurol. Soc. 64 (3), 255–265.

Fox, N., Warrington, E., Freeborough, P., Hartikainen, P., Kennedy, A., Stevens, J., Rossor, M.N., 1996. Presymptomatic hippocampal atrophy in Alzheimer's disease: a longitudinal MRI study. Brain 119 (6), 2001–2007.

Freeborough, P.A., Fox, N.C., 1997. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Trans. Med. Imaging 16 (5), 623–629.

Gaitán, M.I., Shea, C.D., Evangelou, I.E., Stone, R.D., Fenton, K.M., Bielekova, B., Massacesi, L., Reich, D.S., 2011. Evolution of the blood–brain barrier in newly forming multiple sclerosis lesions. Ann. Neurol. 70 (1)2011, 22–29.

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J.C., Beltran, B., Ramió-Torrentà, L., Rovira, À., Lladó, X., 2014. A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. Neuroradiology 56 (5), 363–374.

Gao, Y., Prastawa, M., Styner, M., Piven, J. and Gerig, G. (2014), A joint framework for 4D segmentation and estimation of smooth temporal appearance changes, in 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 1291–1294.

Gerig, G., Welti, D., Guttmann, C.R.G., Colchester, A.C.F., Székely, G., 2000. Exploring the discrimination power of the time domain for segmentation and characterization of active lesions in serial MR data. Med. Image Anal. 4 (1), 31–42.

Giedd, J.N., Blumenthal, J., Jeffries, N.O., Castellanos, F.X., Liu, H., Zijdenbos, A., Paus, T., Evans, A.C., Rapoport, J.L., 1999. Brain development during childhood and adolescence: a longitudinal MRI study. Nat. Neurosci. 2 (10), 861–863.

Group, B.D.C. (2012), Total and regional brain volumes in a population-based normative sample from 4 to 18 years: the NIH MRI Study of Normal Brain Development, Cerebral Cortex 22(1), 1–12.

Guizard, N., Fonov, V.S., García-Lorenzo, D., Nakamura, K., Aubert-Broche, B., Collins, D.L., 2015. Spatio-Temporal regularization for longitudinal registration to subject-specific 3D template. PloS one 10 (8), e0133352.

Hajnal, J.V., Saeed, N., Oatridge, A., Williams, E.J., Young, I.R., Bydder, G.M., 1995. Detection of subtle brain changes using subvoxel registration and subtraction of serial MR images. J. Comput. Assisted Tomogr. 19 (5), 677–691.

Halliday, G., 2017. Pathology and hippocampal atrophy in Alzheimer's disease. Lancet Neurol. 16 (11), 862–864.

Holland, D., Dale, A.M., 2011. Nonlinear registration of longitudinal images and measurement of change in regions of interest. Med. Image Anal. 15 (4), 489–497.

Iglesias, J.E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., Reuter, M., 2016. Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. NeuroImage 141, 542–555.

Klistorner, S., Barnett, M.H., Yiannikas, C., Barton, J., Parratt, J., You, Y., Graham, S.L., Klistorner, A., 2021. Expansion of chronic lesions is linked to disease progression in relapsing–remitting multiple sclerosis patients. Multiple Sclerosis J. 27 (10), 1533–1542.

Krüger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H.H., Schlaefer, A., Schippling, S., 2020. Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. NeuroImage: Clinical 28, 102445.

Laakso, M.P., Soininen, H., Partanen, K., Lehtovirta, M., Hallikainen, M., Hänninen, T., Helkala, E.-L., Vainio, P., Riekkinen, P.J., 1998. MRI of the Hippocampus in Alzheimer's Disease: Sensitivity, Specificity, and Analysis of the Incorrectly Classified Subjects. Neurobiol. Aging 19 (1), 23–31.

Lemieux, L., Wieshmann, U.C., Moran, N.F., Fish, D.R., Shorvon, S.D., 1998. The detection and significance of subtle changes in mixed-signal brain lesions by serial mri scan matching and spatial normalization. Medical Image Anal. 2 (3), 227–242.

Li, Y., Wang, Y., Xue, Z., Shi, F., Lin, W., Shen, D., Initiative, A.D.N., et al., 2010. Consistent 4D cortical thickness measurement for longitudinal neuroimaging study. International Conference on Medical Image Computing and Computer-Assisted Intervention 133–142.

Lombardi, G., Crescioli, G., Cavedo, E., Lucenteforte, E., Casazza, G., Bellatorre, A.-G., Lista, C., Costantino, G., Frisoni, G., Virgili, G., et al., 2020. Structural magnetic resonance imaging for the early diagnosis of dementia due to Alzheimer's disease in people with mild cognitive impairment. Cochrane Database of Systematic Reviews 3 (3).

Lorscheider, J., Buzzard, K., Jokubaitis, V., Spelman, T., Havrdova, E., Horakova, D., Trojano, M., Izquierdo, G., Girard, M., Duquette, P., Prat, A., Lugaresi, A., Grand'Maison, F., Grammond, P., Hupperts, R., Alroughani, R., Sola, P., Boz, C., Pucci, E., Lechner-Scott, J., Bergamaschi, R., Oreja-Guevara, C., Iuliano, G., Van Pesch, V., Granella, F., Ramo-Tello, C., Spitaleri, D., Petersen, T., Slee, M., Verheul, F., Ampapa, R., Amato, M.P., McCombe, P., Vucic, S., Sánchez Menoyo, J.L., Cristiano, E., Barnett, M.H., Hodgkinson, S., Olascoaga, J., Saladino, M.L., Gray, O., Shaw, C., Moore, F., Butzkueven, H., Kalincik, T. and on behalf of the MSBase Study Group (2016), Defining secondary progressive multiple sclerosis, Brain 139(9), 2395–2405.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J. M., 2013. MIRIAD–Public release of a multiple time point Alzheimer's MR imaging dataset. NeuroImage 70, 33–36.

Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults'. J. Cognit. Neurosci. 22 (12), 2677–2684.

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Wiestler, B., Berger, C., Eichinger, P.,

Muhlau, M., Reyes, M., Salmen, A., Chan, A., Wiest, R. and Wagner, F. (2020), Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence, NeuroImage: Clinical 25.

Metcalf, D., Kikinis, R., Guttmann, C., Vaina, L. and Jolesz, F. (1992), 4d connected component labelling applied to quantitative analysis of ms lesion temporal development, in 1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 3, pp. 945–946.

Mills, K.L., Siegmund, K.D., Tamnes, C.K., Ferschmann, L., Wierenga, L.M., Bos, M.G., Luna, B., Li, C., Herting, M.M., 2021. Inter-individual variability in structural brain development from late childhood to young adulthood. Neuroimage 242, 118450.

Nakamura, K., Fox, R., Fisher, E., 2011. CLADA: cortical longitudinal atrophy detection algorithm. Neuroimage 54 (1), 278–289.

Nakamura, K., Guizard, N., Fonov, V.S., Narayanan, S., Collins, D.L., Arnold, D.L., 2014. Jacobian integration method increases the statistical power to measure gray matter atrophy in multiple sclerosis. NeuroImage: Clinical 4, 10–17.

Oh, J., Ontaneda, D., Azevedo, C., Klawiter, E.C., Absinta, M., Arnold, D.L., Bakshi, R., Calabresi, P.A., Crainiceanu, C., Dewey, B., et al., 2019. Imaging outcome measures of neuroprotection and repair in MS: a consensus statement from NAIMS. Neurology 92 (11), 519–533.

Pirko, I. and Johnson, A. (2008), Neuroimaging of demyelination and remyelination models, Advances in multiple Sclerosis and Experimental Demyelinating Diseases pp. 241–266.

Pongratz, V., Schmidt, P., Bussas, M., Grahl, S., Gaser, C., Berthele, A., Hoshi, M.M., Kirschke, J., Zimmer, C., Hemmer, B., Mühlau, M., 2019. Prognostic value of white matter lesion shrinking in early multiple sclerosis: An intuitive or naïve notion? Brain Behav. 9 (12).

Prastawa, M., Awate, S.P., Gerig, G., 2012. Building spatiotemporal anatomical models using joint 4-D segmentation, registration, and subject-specific atlas estimation. 2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, IEEE 49–56.

Preziosa, P., Pagani, E., Meani, A., Moiola, L., Rodegher, M., Filippi, M., Rocca, M.A., 2022. Slowly Expanding Lesions Predict 9-Year Multiple Sclerosis Disease Progression. Neurol.-Neuroimmunol. Neuroinflamm. 9 (2).

Puonti, O., Iglesias, J.E., Van Leemput, K., 2016. Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. NeuroImage 143, 235–249.

Reuter, M., Fischl, B., 2011. Avoiding asymmetry-induced bias in longitudinal image processing. NeuroImage 57 (1), 19–21.

Reuter, M., Schmansky, N.J., Rosas, D.H., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61 (4), 1402–1418.

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.

Scahill, R.I., Frost, C., Jenkins, R., Whitwell, J.L., Rossor, M.N., Fox, N.C., 2003. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. Arch. Neurol. 60 (7), 989–994.

Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., Bellenberg, B., Zipp, F., Groppa, S., Sämann, P.G., Weber, F., Gaser, C., Franke, T., Bussas, M., Kirschke, J., Zimmer, C., Hemmer, B. and Mühlau, M. (2019), Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging, NeuroImage: Clinical 23.

Sederevičius, D., Vidal-Piñeiro, D., Sørensen, Ø., van Leemput, K., Iglesias, J.E., Dalca, A. V., Greve, D.N., Fischl, B., Bjørnerud, A., Walhovd, K.B., et al., 2021. Reliability and sensitivity of two whole-brain segmentation approaches included in FreeSurfer–ASEG and SAMSEG. NeuroImage 237, 118113.

Sepahvand, N.M., Arnold, D.L. and Arbel, T. (2020), Cnn detection of new and enlarging multiple sclerosis lesions from longitudinal mri using subtraction images, in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 127–130.

Sethi, V., Nair, G., Absinta, M., Sati, P., Venkataraman, A., Ohayon, J., Wu, T., Yang, K., Shea, C., Dewey, B.E., et al., 2017. Slowly eroding lesions in multiple sclerosis. Multiple Sclerosis J. 23 (3), 464–472.

Shi, F., Fan, Y., Tang, S., Gilmore, J.H., Lin, W., Shen, D., 2010. Neonatal brain image segmentation in longitudinal MRI studies. NeuroImage 49 (1), 391–400.

Shi, F., Yap, P.T., Gilmore, J.H., Lin, W. and Shen, D. (2010), Spatial-temporal constraint for segmentation of serial infant brain MR images, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 6326, pp. 42–50.

Smith, S.M., De Stefano, N., Jenkinson, M., Matthews, P.M., 2001. Normalized accurate measurement of longitudinal brain change. J. Comput. Assist. Tomogr. 25 (3), 466–475.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., Stefano, D., 2002. Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis. NeuroImage 17 (1), 479–489.

Solomon, J. and Sood, A. (2004), 4-d lesion detection using expectation-maximization and hidden markov model, in 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro, Vol. 1, pp. 125–128.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H.-H., Jewells, V., Warfield, S., 2008. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. MIDAS J. 1–6.

Sweeney, E., Shinohara, R., Shea, C., Reich, D., Crainiceanu, C.M., 2013. Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. Am. J. Neuroradiol. 34 (1), 68–73.

Tamnes, C.K., Walhovd, K.B., Dale, A.M., Østby, Y., Grydeland, H., Richardson, G., Westlye, L.T., Roddey, J.C., Hagler Jr, D.J., Due-Tønnessen, P., et al., 2013. Brain development and aging: overlapping and unique patterns of change. Neuroimage 68, 63–74.

Thirion, J., Calmon, G., 1999. Deformation analysis to detect and quantify active lesions in three-dimensional medical image sequences. IEEE Trans. Med. Imaging 18 (5), 429–441.

Tustison, N.J., Holbrook, A.J., Avants, B.B., Roberts, J.M., Cook, P.A., Reagh, Z.M., Duda, J.T., Stone, J.R., Gillen, D.L., Yassa, M.A., et al., 2019. Longitudinal mapping of cortical thickness measurements: An Alzheimer's Disease Neuroimaging Initiative-based evaluation study. J. Alzheimer's Disease 71 (1), 165–183.

Van Leemput, K., 2009. Encoding probabilistic brain atlases using Bayesian inference. IEEE Trans. Med. Imaging 28 (6), 822–837.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based bias field correction of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 885–896.

Wang, L., Shi, F., Li, G., Shen, D., 2013. 4D Segmentation of Brain MR Images with Constrained Cortical Thickness Variation. PLOS ONE 8 (7), 1–13.

Wang, L., Shi, F., Yap, P.-T., Gilmore, J.H., Lin, W., Shen, D., 2011. Accurate and Consistent 4D Segmentation of Serial Infant Brain MR Images. Multimodal Brain Image Analysis 7012, 93–101.

Wei, J., Shi, F., Cui, Z., Pan, Y., Xia, Y., Shen, D., 2021. Consistent Segmentation of Longitudinal Brain MR Images with Spatio-Temporal Constrained Networks. International Conference on Medical Image Computing and Computer-Assisted Intervention 89–98.

Wells, W.M., Grimson, W.E., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of MRI data. IEEE Trans. Med. Imaging 15 (4), 429–442.

Welti, D., Gerig, G., Radü, E., L, K. and Székely, G. (2001), Spatio-temporal segmentation of active multiple sclerosis lesions in serial MRI data, in Information Processing in Medical Imaging, 17th International Conference, IPMI, pp. 438–445.

Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., Initiative, A.D.N., et al., 2010. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. NeuroImage 52 (1), 109–118.

Xue, Z., Shen, D., Davatzikos, C., 2005. CLASSIC: Consistent Longitudinal Alignment and Segmentation for Serial Image Computing. In: Information Processing in Medical Imaging, vol. 3565. Springer, Berlin Heidelberg, pp. 101–113.

Xue, Z., Shen, D., Davatzikos, C., 2006. CLASSIC: consistent longitudinal alignment and segmentation for serial image computing. NeuroImage 30 (2), 388–399.