NeuroImage: Clinical 2 (2013) 402–413

Contents lists available at SciVerse ScienceDirect

NeuroImage: Clinical

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

OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI☆

Elizabeth M. Sweeney [a,b,*], Russell T. Shinohara [b,c], Navid Shiee [d], Farrah J. Mateen [e], Avni A. Chudgar [f], Jennifer L. Cuzzocreo [g], Peter A. Calabresi [h], Dzung L. Pham [d], Daniel S. Reich [a,b,g,h], Ciprian M. Crainiceanu [a]

[a] Department of Biostatistics, The Johns Hopkins University, Baltimore, MD 21205, USA
[b] Translational Neuroradiology Unit, Neuroimmunology Branch, National Institute of Neurological Disease and Stroke, National Institute of Health, Bethesda, MD 20892, USA
[c] Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[d] Center for Neuroscience and Regenerative Medicine, Henry M. Jackson Foundation, Bethesda, MD 20892, USA
[e] Department of International Health, The Johns Hopkins University, Baltimore, MD 21205, USA
[f] Department of Radiology, Division of Emergency Radiology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[g] Department of Radiology, The Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA
[h] Department of Neurology, The Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

ARTICLE INFO

Article history:
Received 18 January 2013
Received in revised form 23 February 2013
Accepted 5 March 2013
Available online xxxx

Keywords:
Multiple sclerosis
MRI
Brain
Lesion segmentation
Statistical modeling
Logistic regression

ABSTRACT

Magnetic resonance imaging (MRI) can be used to detect lesions in the brains of multiple sclerosis (MS) patients and is essential for diagnosing the disease and monitoring its progression. In practice, lesion load is often quantified by either manual or semi-automated segmentation of MRI, which is time-consuming, costly, and associated with large inter- and intra-observer variability. We propose OASIS is Automated Statistical Inference for Segmentation (OASIS), an automated statistical method for segmenting MS lesions in MRI studies. We use logistic regression models incorporating multiple MRI modalities to estimate voxel-level probabilities of lesion presence. Intensity-normalized T1-weighted, T2-weighted, fluid-attenuated inversion recovery and proton density volumes from 131 MRI studies (98 MS subjects, 33 healthy subjects) with manual lesion segmentations were used to train and validate our model. Within this set, OASIS detected lesions with a partial area under the receiver operating characteristic curve for clinically relevant false positive rates of 1% and below of 0.59% (95% CI; [0.50%, 0.67%]) at the voxel level. An experienced MS neuroradiologist compared these segmentations to those produced by LesionTOADS, an image segmentation software that provides segmentation of both lesions and normal brain structures. For lesions, OASIS out-performed LesionTOADS in 74% (95% CI: [65%, 82%]) of cases for the 98 MS subjects.

To further validate the method, we applied OASIS to 169 MRI studies acquired at a separate center. The neuroradiologist again compared the OASIS segmentations to those from LesionTOADS. For lesions, OASIS ranked higher than LesionTOADS in 77% (95% CI: [71%, 83%]) of cases. For a randomly selected subset of 50 of these studies, one additional radiologist and one neurologist also scored the images. Within this set, the neuroradiologist ranked OASIS higher than LesionTOADS in 76% (95% CI: [64%, 88%]) of cases, the neurologist 66% (95% CI: [52%, 78%]) and the radiologist 52% (95% CI: [38%, 66%]).

OASIS obtains the estimated probability for each voxel to be part of a lesion by weighting each imaging modality with coefficient weights. These coefficients are explicit, obtained using standard model fitting techniques, and can be reused in other imaging studies. This fully automated method allows sensitive and specific detection of lesion presence and may be rapidly applied to large collections of images.

1. Introduction

Multiple sclerosis (MS) is an inflammatory disease of the brain and spinal cord characterized by demyelinating lesions that are most easily identified, at least on magnetic resonance imaging (MRI) studies, in the white matter of the brain (Sahraian and Radue, 2007). Quantitative analyses of MRI, such as the number and volume of lesions, are essential for diagnosing the disease and monitoring its progression (Rovira and León, 2008; Rovira et al., 2009). MRI measures are also a common

* Corresponding author at: Department of Biostatisitcs, 615 N Wolfe Street, E3518, Baltimore, MD 21205, USA. Tel.: +1 317 698 5700.
E-mail address: emsweene@jhsph.edu (E.M. Sweeney).

2213-1582/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.nicl.2013.03.002

primary endpoint in phase II immunomodulatory drug therapy trials (Sormani et al., 2009). In these trials, either manual or semi-automated segmentations are used to compute the total number of lesions and the total lesion volume (Lladó et al., 2011). Manual delineation is challenging as three-dimensional information from several MRI modalities must be integrated (Lladó et al., 2011). Manual assessment of MRI is also prone to large inter- and intra-observer variability (Simon et al., 2006). While semi-automated methods have been found to decrease inter- and intra-rater variability, they still require manual reader input and are time consuming (García-Lorenzo et al., 2013). Therefore a sensitive and specific automated method to detect lesions in the brain is essential for the analysis of studies with a high numbers of MS patients.

Lladó et al. (2011) provides a comprehensive review of currently available automated cross-sectional MS lesion segmentation methods, or methods used to identify lesions from a single MRI study. We divide these methods into four categories: supervised classifier with an atlas, supervised classifier with no atlas, unsupervised classifier with an atlas, and unsupervised classifier with no atlas. We focus on supervised methods without atlases, as the method we propose is in this category. Supervised methods without atlases train on manually segmented images annotated by experts and use image intensities of MRI to classify lesions (Lladó et al., 2011). Supervised classification algorithms are applied to the volumes: artificial neural networks (Goldberg-Zimring et al., 1998), spatial clustering (Alfano et al., 2000), k-nearest neighbors (Anbeek et al., 2004, 2005, 2008), Parzen window (Sajja et al., 2006), Parzen window and morphological grayscale reconstruction (Datta et al., 2006), Bayes (Scully et al., 2008), AdaBoost (Morra et al., 2008), simulated annealing and Markov random fields (Subbanna et al., 2009), and graph cuts (Lecoeur et al., 2009). All of the aforementioned methods except Anbeek et al. (2008) use multi-modality MRI information to classify lesions. The most widely-used feature across all segmentation methods is voxel intensity, which derives strength from a multi-modality approach (Lladó et al., 2011).

The method we propose uses a logistic regression model to assign voxel-level probabilities of lesion presence in structural MRI of patients with MS. Logistic regression models have been used for segmentation of brain tissues and pathology in MRI (Bullmore et al., 1999; Dinh et al., 2012; Lee et al., 2005). For applications to MS, logistic regression has been used for detection of gadolinium enhancing lesions (Karimaghaloo et al., 2012), prediction of gadolinium enhancing lesions without administering contrast agents (Shinohara et al., 2012), and for segmentation of new and enlarging MS lesions (Sweeney et al., 2013). To our knowledge logistic regression has not been used in cross-sectional segmentation of MS lesions in structural MRI.

One difficulty in automated segmentation of MRI is due to variable imaging acquisition parameters (Lladó et al., 2011). All of the segmentation methods reviewed in Lladó et al. (2011) have tuning parameters that are adjusted to a particular data set and may not generalize to a new data set with different acquisition parameters. These parameters are not informed by the data and therefore must be tuned empirically, often with little to no interpretability of the parameter. Application to a new data set may require several iterations of segmentations to adjust the tuning parameters to values that produce acceptable segmentations. A method in which the tuning parameters are informed by the data and for which adjustments are intuitive and simple would therefore be valuable.

A second difficulty in intensity-based segmentation is that MRI data are acquired in arbitrary units; units can vary widely between and within imaging centers. These variations are attributed to scanner hardware, interactions between hardware and patients, and variations in acquisition parameters (Simmons et al., 1994). Therefore, proper intensity normalization is essential in developing a generalizable segmentation method. Many of the segmentation methods use intensity-normalized volumes (Lladó et al., 2011), but these methods do not demonstrate the generalizability of the normalization procedure to changes in imaging acquisition parameters and imaging centers. In García-Lorenzo et al. (2013) the authors performed a PubMed and Google Scholar search for MS lesion segmentation papers. Of the 47 papers that met their search criteria, only 13 of these papers used multicenter data for validation, and the largest database used for validation consisted of 41 subjects. To show generalizability, methods must be validated on multicenter data with many subjects.

A third difficulty is intensity inhomogeneity, the slow spatial intensity variations of the same tissue within an MRI volume. Inhomogeneity can significantly reduce the accuracy of image segmentation (Hou, 2006), and therefore some form of spatial normalization is necessary for accurate lesion segmentation. Most lesion segmentation methods assume that these inhomogeneities have been corrected during image preprocessing, but we have found strong spatial patterns within tissue type even after the N3 inhomogeneity correction algorithm (Sled et al., 1998) is applied.

To address these and related problems, we propose OASIS is Automated Statistical Inference for Segmentation (OASIS), a fully automated, generalizable, and novel statistical method for cross-sectional MS lesion segmentation. Using intensity information from multiple modalities of MRI, a logistic regression model assigns voxel-level probabilities of lesion presence. After training on manual segmentations, the OASIS model produces interpretable results in the form of regression coefficients that can be applied to imaging studies quickly and easily. OASIS uses intensity-normalized brain MRI volumes, enabling the model to generalize to changes in scanner and acquisition sequence. OASIS also adjusts for intensity inhomogeneities that preprocessing bias field correction procedures do not remove, using smoothed volumes. This allows for more accurate segmentation of brain areas that are highly distorted by inhomogeneities, such as the cerebellum. One of the most practical properties of OASIS is that the method is fully transparent, easy to explain and implement, and simple to modify for new data sets.

To illustrate the generalizability of OASIS to changes in imaging acquisition parameters, we evaluated the performance of the algorithm on a total of 300 MRI studies from two separate imaging centers with varying acquisition parameters. This is a crucial criterion for assessing the generalizability and utility of the method.

2. Materials and methods

In this section we introduce OASIS, a method inspired by Subtraction Based Inference for Modeling and Estimation (SuBLIME), an automated method for the longitudinal segmentation of incident and enlarging MS lesions (Sweeney et al., 2013). Before the OASIS logistic regression model is fit, a brain tissue mask is created, all MRI volumes are intensity normalized, and smoothed volumes are created to capture local spatial information and adjust for remaining field inhomogeneities. The OASIS method involves two iterations of model fitting: the first to perform an initial lesion segmentation and the second to use this initial lesion segmentation to remove lesions, which can distort the smoothed volumes. After the final model is fit, the regression coefficients are applied to produce three dimensional maps of voxel-level probabilities of lesion presence.

We evaluate the performance of OASIS on MRI volumes of the brain acquired with various acquisition protocols. We use data sets from two different imaging centers for validation, which we refer to as Validation Set 1 and Validation Set 2. Validation Set 1 has manual lesion segmentations. We trained the OASIS method on a subset of the studies in this dataset, and tested on the remaining studies. An expert evaluated the segmentations from Validation Set 1. Validation Set 2 is used to demonstrate generalizability to changes in image acquisition parameters. We applied the coefficients from the model trained on Validation Set 1 to the studies in Validation Set 2, and experts evaluated the OASIS lesion segmentations.

2.1. Study population

Validation Set 1 contains a total of 131 MRI studies from 131 subjects. Of these studies, 98 are from patients with MS and 33 are healthy volunteer scans. Of the 98 patients with MS, the median age is 44 years (IQR: [33, 54]), 72 are female (26 male), and the median EDSS is 3.5 (IQR:[2, 6]). The median age of the healthy volunteers is 34 (IQR: [28, 42]) and 19 are female (14 male).

Validation Set 2 contains a total of 169 MRI studies from 149 subjects. Twenty subjects in Validation Set 2 have baseline and follow-up scans. The mean time between baseline and follow-up for these 20 subjects is 132 days (IQR: [51, 182]). The subjects in the validation set are a mixture of healthy volunteers and patients: 110 of the patients have MS, 38 have other neurological diseases, and one is a healthy volunteer. The median age of the MS patients is 42 (IQR: [33,50]); 54 are female (56 male); 68 have relapsing remitting MS, 31 have primary progressive MS, and 11 have secondary progressive MS. The median age of the patients with other neurological diseases is 41 years, (IQR: [35, 51]) and 8 are female (30 male). The healthy volunteer is a 28 year old female.

2.2. Experimental methods

T1-weighted, T2-weighted, fluid-attenuated inversion recovery (FLAIR) and proton density (PD) volumes were acquired for all subjects at each study, and all imaging protocols were approved by local institutional review boards. For Validation Set 1, 3DT1-MPRAGE images (repetition time (TR) = 10 ms; echo time (TE) = 6 ms; flip angle (FA) $\alpha = 8°$; inversion time (TI) = 835 ms, resolution = 1.1 mm × 1.1 mm × 1.1 mm), 2D T2-weighted pre-contrast FLAIR images (TR = 11,000 ms; TE = 68 ms; TI = 2800 ms; in-plane resolution = 0.83 mm × 0.83 mm; slice thickness = 2.2 mm), T2-weighted and PD images (TR = 4200 ms; TE = 12/80 ms; resolution = 0.83 mm × 0.83 mm × 2.2 mm) were acquired on a 3 T MRI scanner (Philips Medical Systems, Best, The Netherlands).

For Validation Set 2, the 3D T2-weighted post-contrast FLAIR was acquired using a variable flip angle sequence, the 2D PD and T2-weighted volumes using a dual-echo fast-spin-echo sequence, and the 3D T1-weighted volume using an inversion-prepared fast spoiled gradient-echo sequence. These studies were acquired on a single 3 T MRI scanner (Signa Excite HDxt; GE Healthcare, Milwaukee, Wisconsin). Table 1 contains the ranges for the Validation Set 2 scanning parameters.

2.3. Image preprocessing

Before building our statistical model for the lesion segmentation, we preprocessed the images from Validation Set 1 and Validation Set 2 using the tools provided in Medical Image Processing Analysis and Visualization (MIPAV) (McAuliffe et al., 2001), TOADS-CRUISE (http://www.nitrc.org/projects/toads-cruise/), and Java Image Science Toolkit (JIST) (Lucas et al., 2010) software packages. We first rigidly aligned the T1-weighted image of each subject into the Montreal Neurological Institute (MNI) standard space (voxel resolution1mm^3). We then registered the FLAIR, PD, and T2-weighted images of each subject to the aligned T1-weighted images. We also applied the N3 inhomogeneity correction algorithm (Sled et al., 1998) to all images and removed extracerebral voxels using SPECTRE, a skull-stripping procedure (Carass et al., 2011).

Table 1
Ranges for Validation Set 2 scanning parameters.

	FA (degrees)	TR (ms)	TE (ms)	TI (ms)
FLAIR	90	(4800, 8802)	(124.3, 151.4)	(1481, 2200)
T2-weighted	90	5317	(116.2, 124.2)	NA
PD	90	5317	(16.0, 23.7)	NA
T1-weighted	(6, 13)	(8.7, 9.1)	(3.2, 3.6)	(450, 725)

2.4. Statistical modeling and spatial smoothing

We performed all statistical modeling in the R environment (version 2.12.0, R Foundation for Statistical Computing, Vienna, Austria) with the packages AnalyzeFMRI (Bordier et al., 2009), biglm (Lumley, 2009), ff (Adler et al., 2011), and ROCR (Sing et al., 2009). We used the FSL tool fslmaths (http://www.fmrib.ox.ac.uk/fsl) for the three dimensional spatial smoothing of the volumes.

2.5. Brain tissue mask

The first step in OASIS is to create a mask of the brain that excludes cerebrospinal fluid (CSF). CSF is excluded because it disrupts the capture of the inhomogeneity field and distorts the representation of the local cerebral features when creating smoothed volumes. To make this mask, we used the extracerebral voxel removal mask described in the Image Preprocessing Section and excluded voxels in the mask that appear hypointense in the FLAIR volume. Because CSF is hypointense in the FLAIR, we empirically found that excluding voxels falling below the bottom 15th percentile of FLAIR intensities over the extracerebral voxel removal mask removes CSF outside of the brain and in the ventricles. We refer to this mask as the brain tissue mask. Fig. 2B1 shows a slice of the brain tissue mask for a particular subject for illustration.

2.6. Intensity normalization

We used intensities from the FLAIR, PD, T2-weighted, and T1-weighted volumes to identify the presence of MS lesions. We denote the observed intensity of voxel v, for subject i by:

$$M_i^0(v), M = FLAIR, PD, T2, T1$$

where M indicates the imaging sequence.

MRI volumes are acquired in arbitrary units. Analyzing images across subjects and imaging centers requires that images be normalized so that voxel intensities have common interpretations. For normalization, we adapt the normalization method from Shinohara et al. (2011) to normalization with respect to the brain tissue mask. The normalized intensity of voxel v, for subject i is denoted by:

$$M_i^N(v) = \frac{M_i^0(v) - \mu_{i,M}^0}{\sigma_{i,M}^0}$$

where $\mu_{i,M}^0$ and $\sigma_{i,M}^0$ are the mean and standard deviation of the observed voxel intensities in the brain tissue mask of subject i, from sequence M. The normalized voxel intensities are standard scores of the brain tissue mask. Fig. 1A shows a slice of the normalized images from all four modalities from a single subject with MS: FLAIR, T2-weighted, PD, and T1-weighted.

2.7. Smoothed volumes

To account for intensity inhomogeneities that remain after initial inhomogeneity correction, we use a sequence of multiresolution smoothed volumes, obtained using different levels of smoothing. The smoothed volumes are created by three dimensional smoothing of the normalized volume from each modality over the brain tissue mask. A Gaussian smoother with relatively large kernel window size is used to smooth over the features in the brain and capture the pattern of the remaining inhomogeneity.

For subject i and imaging modality M, let k be the size of the kernel window. Then the intensity in voxel v of the smoothed volume of

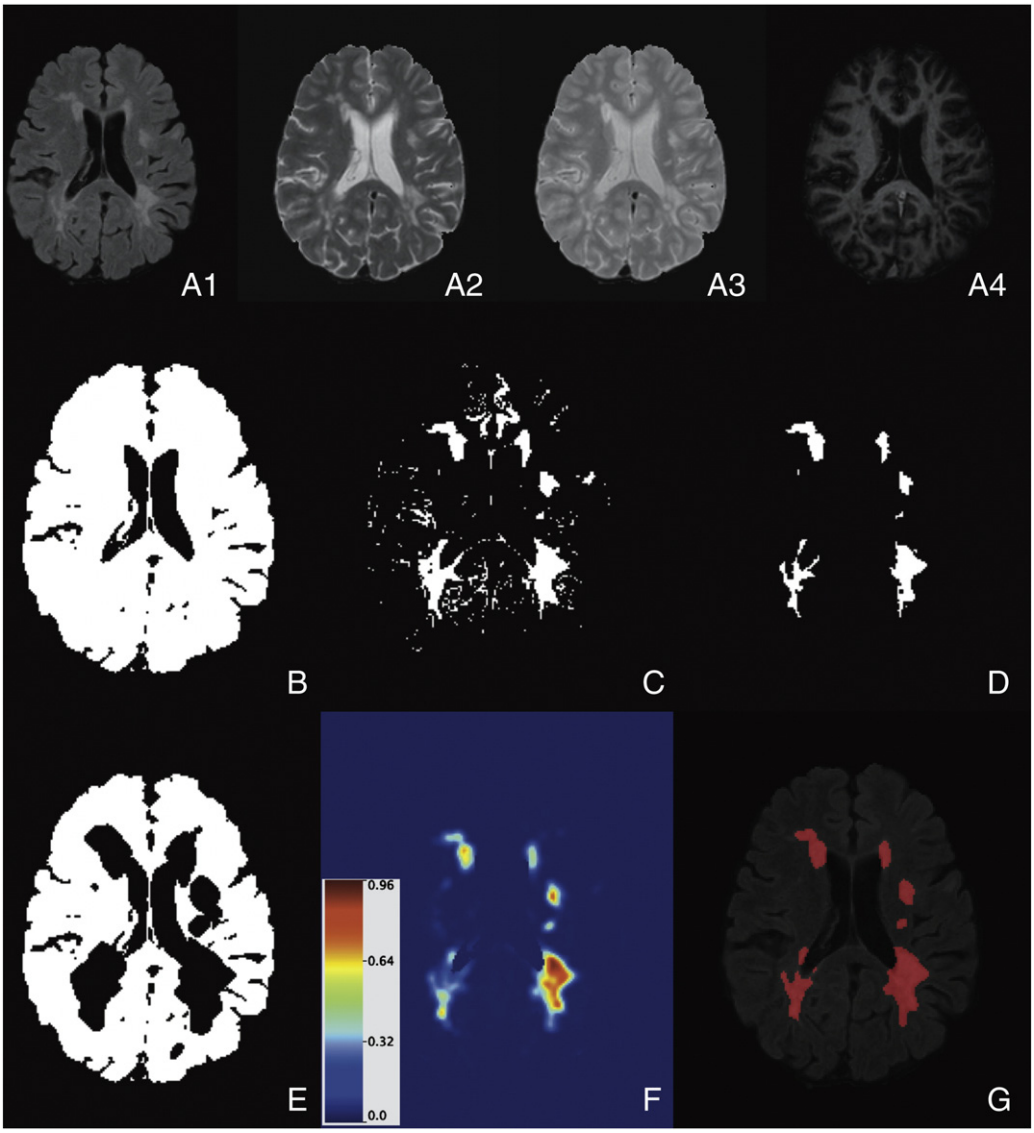

Fig. 1. A. Axial slice from different modalities of intensity normalized brain MRI of a single subject: A1. FLAIR image. A2. T2-weighted image. A3. PD image. A4. T1-weighted image. B. Brain tissue mask of an axial slice of the brain. C. Axial slice of select voxels for OASIS modeling. D. Manual lesion segmentation of an axial slice of the brain. E. Axial slice of brain tissue mask with dilated lesion mask made at a false positive rate of 1% removed. F. Axial slice of the smoothed probability map with intensity scale. G. Binary segmentation of the probability map from the OASIS model at false positive rate of .005 overlaid on the FLAIR image. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

imaging modality M is expressed as $\mathcal{G}M_i^N(v,k)$. The smoothed volumes are used in the OASIS model to incorporate spatial information and to account for inhomogeneities in the brain that persist after N3 correction. For OASIS we use smoothed volumes as covariates with kernel window sizes of 10 and 20 voxels, which were found empirically on Validation Set 1 to work well. Fig. 2 shows the smoothed volumes for both kernel window sizes of 10 and 20 from each modality. The kernel window size of 20 smooths over the anatomical features almost completely, while the kernel window size of 10 still preserves some of these features, such as the hyperintensities of the gray matter in the FLAIR, T2-weighted, and PD volumes and hypointensities of the gray matter in the T1-weighted volume.

2.8. OASIS is Automated Statistical Inference for Segmentation

In this section we introduce the OASIS model. OASIS uses logistic regression to model the probability that a voxel is part of a lesion. We choose logistic regression because it is extremely simple and easy to interpret. We model lesions at the voxel level using FLAIR, PD, T2-weighted, and T1-weighted intensities as well as the intensities from the smoothed volumes of each modality with kernel window sizes of 10 and 20 voxels. The model must be trained on a gold standard measure of lesion presence. Fig. 1D is an example of manual lesion segmentation, which is an appropriate gold standard measure for the OASIS model. The result of our model is a collection of coefficients that can be used to create three-dimensional maps of the probabilities of lesion presence. OASIS obtains the estimated logit of the probability of each voxel being part of a lesion by weighting these 12 images (the four imaging modalities and smoothed volumes for each modality) with the coefficients.

The first step of the modeling procedure is to select candidate voxels to minimize false positives and computation time. Lesions appear as hyperintensities in the FLAIR volume. The brain tissue mask was applied to the FLAIR volume, and the 85th percentile and above of voxels in the brain tissue mask were selected as candidate voxels for lesion presence. In Validation Set 1, there were a total of 1,093,394 lesion voxels (a

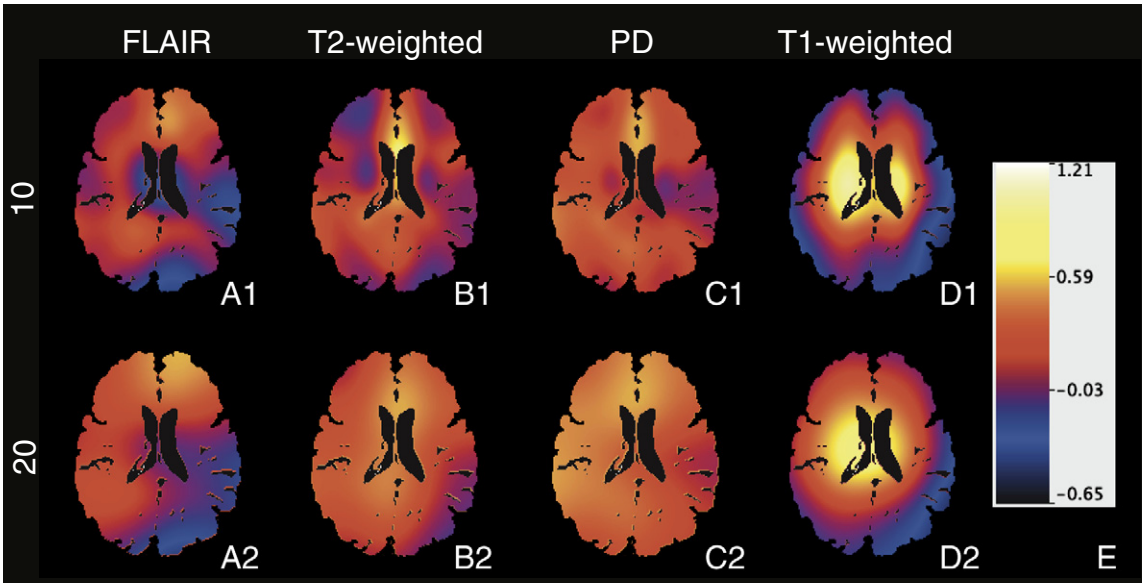

Fig. 2. Axial slice from a single subject of the smoothed volumes from all modalities. Row one contains the smoothed volumes with kernel window size of 10 and row two contains the smoothed volumes with kernel window size of 20. Column A contains the FLAIR images, B contains the T2-weighted images, C contains the PD images and D contains the T1-weighted images. To link the figure with the notation used in this paper: A1. $\mathcal{G}FLAIR_i^N(v,10)$; A2. $\mathcal{G}FLAIR_i^N(v,20)$; B1. $\mathcal{G}T2_i^N(v,10)$; B2. $\mathcal{G}T2_i^N(v,20)$; C1. $\mathcal{G}PD_i^N(v,10)$; C2. $\mathcal{G}PD_i^N(v,20)$; D1. $\mathcal{G}T1_i^N(v,10)$; D2. $\mathcal{G}T1_i^N(v,20)$; and E. Scale of intensities in the smoothed volumes.

volume of 1093 cm^3). The voxel selection procedure excluded 64,556 (6%) of these voxels, but lowered the searchable area to 15% of the original size. This procedure also decreases the number of potential false positive voxels. Using this threshold also significantly decreases the number of voxels the model must be fit on, allowing for a much faster fit. Fig. 1C shows a slice of the voxel selection mask for a single subject.

We then fit a voxel-level logistic regression model over the candidate voxels. In the OASIS model, the probability that a voxel is part of a lesion is represented as $P\{L_i(v)=1\}$, where L is a random variable denoting voxel-level lesion presence. If there is a lesion in voxel v for subject i, then $L_i(v)=1$. Otherwise, $L_i(v)=0$. The probability that a voxel v contains lesion incidence is modeled with the following logistic regression model:

$$\begin{aligned} logit[P\{L_i(v)=1\}] = {} & \beta_0 + \\ & \beta_1 FLAIR_i^N(v) + \beta_2 \mathcal{G}FLAIR_i^N(v,10) + \beta_3 \mathcal{G}FLAIR_i^N(v,20) + \\ & \beta_4 PD_i^N(v) + \beta_5 \mathcal{G}PD_i^N(v,10) + \beta_6 \mathcal{G}PD_i^N(v,20) + \\ & \beta_7 T2_i^N(v) + \beta_8 \mathcal{G}T2_i^N(v,10) + \beta_9 \mathcal{G}T2_i^N(v,20) + \\ & \beta_{10} T1_i^N(v) + \beta_{11} \mathcal{G}T1_i^N(v,10) + \beta_{12} \mathcal{G}T1_i^N(v,20) \\ & + \beta_{13} FLAIR_i^N(v) * \mathcal{G}FLAIR_i^N(v,10) + \beta_{14} FLAIR_i^N(v) * \mathcal{G}FLAIR_i^N(v,20) \\ & + \beta_{15} PD_i^N(v) * \mathcal{G}PD_i^N(v,10) + \beta_{16} PD_i^N(v) * \mathcal{G}PD_i^N(v,20) \\ & + \beta_{17} T2_i^N(v) * \mathcal{G}T2_i^N(v,10) + \beta_{18} T2_i^N(v) * \mathcal{G}T2_i^N(v,20) \\ & + \beta_{19} T1_i^N(v) * \mathcal{G}T1_i^N(v,10) + \beta_{20} T1_i^N(v) * \mathcal{G}T1_i^N(v,20) \end{aligned} \quad [1].$$

The effect of magnetic field inhomogeneities is thought to be multiplicative, so we use the interactions between the normalized volume and the smoothed volume in the model.

2.9. OASIS model refinement

The second iteration of the OASIS model fitting is done to reduce the influence of lesions in the smoothed volumes. First, we fit the model and use the estimated coefficients to create maps of the estimated probability of lesion presence at each voxel. To incorporate spatial information of the neighboring voxels and reduce noise, we smooth the estimated probabilities from the model using a Gaussian kernel with window size of 3 mm. This kernel size was empirically chosen and found to perform well. The resulting probability maps were then thresholded using a liberal false positive rate of 1% (threshold value of 0.10), which resulted in model based hard segmentations of lesions. These lesion masks were then dilated by 5 voxels to ensure that the entire lesion was captured and removed from the brain tissue mask. Fig. 1E shows the brain tissue mask with the lesions removed. New smoothed volumes were created by applying a Gaussian smoother with kernel window sizes of 10 and 20 to the normalized image from each modality over the brain tissue mask with the lesions removed. We inpainted the smoothed volumes to fill the places where lesions were removed with the values we would expect in this area if it were occupied by normal, healthy tissue.

The intensity in voxel v of the normalized image after the second Gaussian smoother has been applied is labeled as, $\mathcal{G}^2 M_i^N(v,k)$. Fig. 3 shows an axial slice for a subject of the FLAIR volume and the smoothed volume for this image with kernel window sizes of 10 and 20 before and after the lesions were removed. To link the figure with the notation, Fig. 3A shows $FLAIR_i^N(v)$, Fig. 3B shows a scale of intensities in the smoothed volumes, Fig. 3C1 shows $\mathcal{G}FLAIR_i^N(v,10)$, Fig. 3C2 shows $\mathcal{G}^2 FLAIR_i^N(v,10)$, Fig. 3D1 shows $\mathcal{G}FLAIR_i^N(v,20)$, and Fig. 3D2 shows $\mathcal{G}^2 FLAIR_i^N(v,20)$. The lesions are captured in the first smoothed volume, especially with the kernel size of 10, but are not captured in the second smoothed volume. The model [1] was refit over the same voxels using the second smoothed volume to obtain the final coefficients that are used to create the final probability maps. Again, the final estimated probabilities are smoothed using a Gaussian kernel with window size of 3 mm. Fig. 1F shows a slice of the probability map for a subject and a scale of intensities. Red indicates areas with a higher probability of being a lesion and blue indicates areas with a lower probability of being a lesion.

2.10. Probability map and binary segmentation

Using this fitted model to generate a probability map for the entire brain from a set of new images takes about 30 min for each study using a standard workstation. The Gaussian smoothing is the slowest step of the algorithm and takes approximately one minute for each volume. These computations can be parallelized to take substantially less time; the entire algorithm can be run in approximately 5 min with 8 cores. To make a probability map for a new study, the two sets of regression coefficients, a brain mask, and the FLAIR, PD, T2-weighted, and T1-weighted volumes are required. Using population-level thresholds, the probability maps from OASIS can be used to create

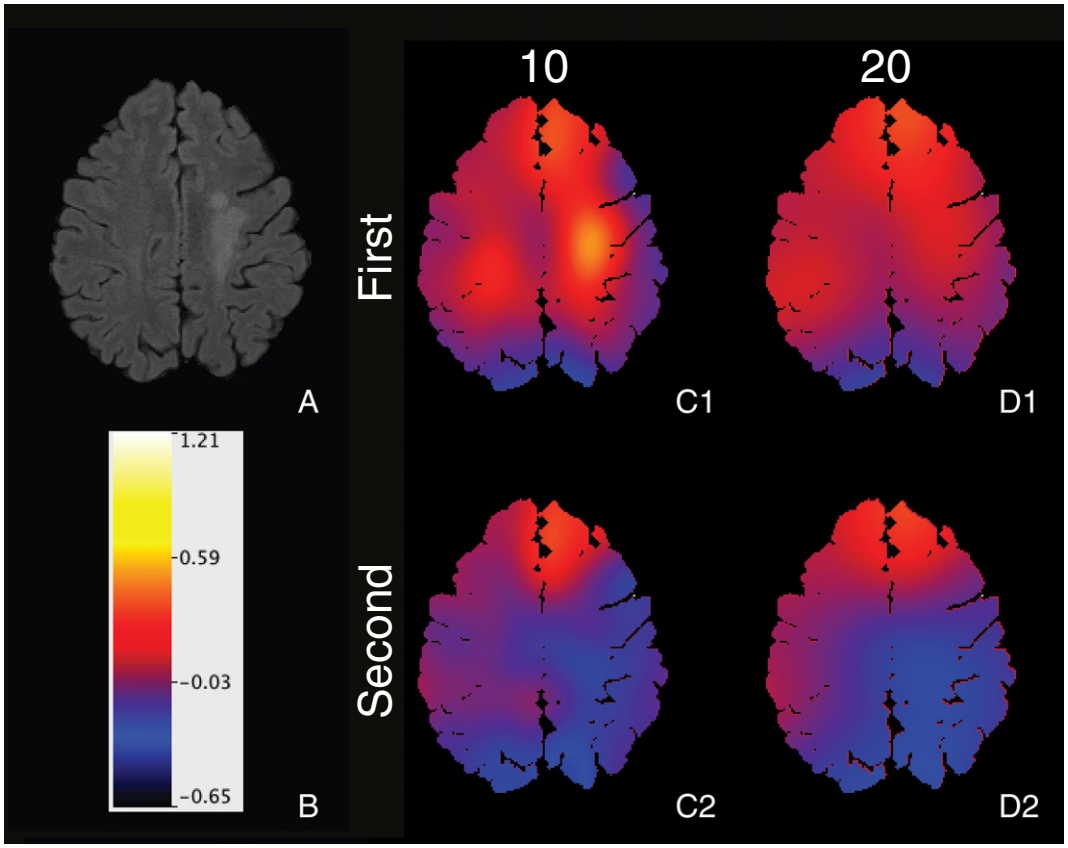

Fig. 3. Axial slice of the FLAIR volume and the first and second smoothed volumes created from the FLAIR image for a single subject. To link the figure with the notation used in this paper: A. $FLAIR_i^N(v)$ B. Scale of intensities in the smoothed volumes C1. $\mathcal{G}FLAIR_i^N(v,10)$; C2. $\mathcal{G}^2FLAIR_i^N(v,10)$; D1. $\mathcal{G}FLAIR_i^N(v,20)$; and D2. $\mathcal{G}^2FLAIR_i^N(v,20)$.

hard segmentations of lesion presence. Fig. 1G shows a slice of a hard segmentation overlaid on the FLAIR image. A summary of how to apply the OASIS method to a new MRI study can be found in the Appendix.

2.11. Validation with gold standard: Validation Set 1

Validation Set 1, described in detail in the Materials and methods section, consists of 131 MRI studies: 98 studies from MS subjects and 33 studies from healthy subjects. To fit the model and to measure performance, we required a set of data in which the outcome is assessed using a gold standard measure. The gold standard was obtained using manual segmentation by a technologist with more than 10 years of experience in delineating white matter lesions. The technologist spent between 30 min to an hour segmenting each study, depending on the lesion load and distribution. The majority of the studies had at least moderate pathology and therefore took between 45 min to an hour. The segmentations were made from the FLAIR and T1-weighted volumes. Fig. 1D shows a manually segmented slice for a subject. The mean volume of lesions for MS subjects in Validation Set 1 is 11.2 cm^3 (IQR: [1.7 cm^3, 16.6 cm^3]). It was assumed that the healthy subjects did not have any lesions.

To evaluate the performance of our model within Validation Set 1, we trained the model [1] on 20 randomly selected subjects (15 MS subjects and 5 healthy subjects) and tested on the remaining 111 subjects (83 MS subjects and 28 healthy subjects). We used only the studies from the 111 subjects in this test set to estimate the voxel-level receiver operating characteristic (ROC) curve and area under the curve (AUC). These performance measures are known to be susceptible to instability. To account for this, we nonparametrically bootstrapped with replacement the subjects to the training and testing sets. We then fit the model on the training set and observed the performance of the model in the testing set.

It is known that the full AUC summarizes test performance over regions of the ROC space that are not clinically relevant for lesion segmentation (Sweeney et al., 2013). Once a test has been able to distinguish well between disease and not disease, the performance of the test for particular applications must be evaluated, in which case one may be interested in only a small portion of the ROC curve (Obuchowski, 2003). In this particular application we are interested in using the lesion segmentation to identify lesions and to provide accurate estimations of lesion volumes. The mean lesion volume of manual lesion segmentations from Validation Set 1 is 11.2 cm^3 (IQR [1.7 cm^3, 16.6 cm^3]). For the entire brain, a false positive rate of .01 would correspond to a volume of 12.8 cm^3 of healthy brain being falsely identified as lesion, which is more than the mean lesion volume in Validation Set 1. Therefore we examined only false positive rates below 1%. We provide the partial ROC curve with bootstrapped 95% confidence bands for clinically relevant false positive rates of 1% and below.

2.12. Validation with expert rankings: Validation Set 1 and Validation Set 2

For the studies in Validation Set 2, gold standard segmentations were not available. To evaluate the performance of OASIS on Validation Set 2, three experts (a neuroradiologist, neurologist, and radiologist) compared OASIS segmentations to those from LesionTOADS, an open-source lesion segmentation software (http://www.nitrc.org/projects/toads-cruise/), (Shiee et al., 2008a,b, 2010). Validation Set 2, described in detail in the Materials and methods section, consists of 169 MRI studies of 149 subjects, 20 of whom had follow-up visits. These studies were acquired using a variety of imaging protocols.

For the OASIS algorithm, the only parameter that must be tuned when moving to a new dataset is the population-level threshold. For Validation Set 2 we used the coefficients that were trained on Validation Set 1 and then empirically adjusted the population level threshold for Validation Set 2. To adjust this threshold, we randomly sampled 10 subjects from Validation Set 2. We applied thresholds between 0.10 and 0.50 (by increments of 0.05) to the probability maps, examined the segmentation, and empirically chose a threshold of 0.35 for Validation Set 2. This threshold adjustment is very fast and transparent. We ran the segmentations for the 10 subjects in parallel, and each segmentation took less than 5 min. Next, we thresholded the probability maps at the 9 different thresholds, which took only seconds. Last, we looked through the segmentations and the original images to select the optimal (most reasonable) threshold, which took only about a minute for each subject. The entire process of tuning the threshold took less than an hour and involved only 10 min of manual image examination. This procedure only needs to be performed once when moving to a new imaging center or study. For the segmentation comparison, we presented the three experts with segmentations at the threshold value of 0.35 on all of the images in Validation Set 2 as well as at the threshold from Validation Set 1 with a false positive rate of 0.005, a threshold value of 0.16. We will refer to the threshold value of 0.35 as the empirically adjusted threshold and the threshold value of 0.16 as the Validation Set 1 threshold.

We compared both OASIS segmentations to the segmentations produced by the open source software LesionTOADS. We ran LesionTOADS with T1-weighted and FLAIR inputs and the default parameters. We adjusted the smoothing parameter from 0.2 to 0.4 because we empirically found this to improve the quality of the segmentations. It is important to note that LesionTOADS not only segments lesions, but also segments the other tissue classes of the brain. For this analysis, we only used the lesion segmentations.

We designed an image rating system to evaluate the performance of the two segmentation algorithms. For each of the 169 studies, we had three segmentations: the LesionTOADS segmentation, the OASIS segmentation with the threshold from Validation Set 1, and the OASIS segmentation with the empirically adjusted threshold. We also randomly selected 20 of the MRI studies and created duplicates of these to assess rating reliability, for a total of 189 studies. We randomized the order in which the segmentations were presented to the experts and randomly assigned each segmentation a letter: A, B, or C, so as to blind the rater to the segmentation algorithm.

We presented each of the 189 MRI studies to an experienced MS neuroradiologist. For each study, the neuroradiologist examined the set of three segmentations along with the original FLAIR, PD, T1-weighted, and T2-weighted volumes. The neuroradiologist then scored the performance of each of the segmentations on a continuous scale from 0 to 100, with 0 being an unusable lesion segmentation and 100 being a perfect segmentation. The neuroradiologist was presented all three segmentations simultaneously, so that scores were assigned relative to one another. Fifty of the studies were selected to be scored with the same system by a neurologist with a subspecialty in MS and a general radiologist in order to assess rater agreement among the three raters. The 50 studies were comprised of 45 randomly selected studies with 5 of the studies repeated to assess rater reliability.

The neuroradiologist also compared and scored the OASIS and LesionTOADS segmentations from the studies for the 98 MS patients in Validation Set 1. This allows for comparison of the performance of the segmentations on Validation Set 1 and Validation Set 2.

3. Experimental results

3.1. Validation Set 1: training with gold standard

The OASIS model has an estimated full AUC of 98% (95% CI; [96%, 99%]) and a partial AUC for clinically relevant false positive rates of 1% and below of 0.59% (95% CI; [0.50%, 0.67%]) in the test set. Fig. 4 shows the voxel-level partial ROC curve for the test set with bootstrapped 95% confidence bands for clinically relevant false positive rates. The probability map threshold that corresponds to a false positive rate of 1% is 0.10. The vertical axis of the partial ROC curve shows the true positive rate (sensitivity) for thresholds between 0 and 0.10 of the probability map and the horizontal axis shows the false positive rate (1 − specificity) for these thresholds.

The coefficients from fitting the logistic model [1] over all 131 studies in Validation Set 1, a total of 24 million voxels, are reported in the Appendix. The coefficients from the first and second fit of the model are provided. We also assessed the variation in the coefficients by nonparametrically bootstrapping the subjects with replacement. The bootstrapped 95% confidence intervals for the coefficients can be found in the Appendix. The variance of these coefficients is large in comparison to the estimates of the coefficients. The instability in the coefficients does not impact the performance of OASIS, as illustrated in the stability of the partial ROC curve.

Choosing a final threshold value after the second probability maps are made is a tradeoff between sensitivity and specificity. OASIS is flexible, and the appropriate false positive rate may be selected for a particular application. Table 2 shows the threshold values, sensitivity, and dice similarity coefficient (Dice, 1945) for four different false positive rates for the model fit over all of the studies in Validation Set 1. OASIS detected lesions in many of the healthy subjects. Table 3 shows the mean volume of false positive lesions detected in the healthy and MS subjects for the four threshold values from Table 2. The volume of false positives for both the MS and healthy subjects is comparable.

3.2. Validation Set 1: neuroradiologist rating results

For the neuroradiologist rankings of the OASIS and LesionTOADS segmentations for the 98 MS subjects in Validation Set 1, we performed a paired *t*-test to assess the difference in the means of the OASIS segmentations and the LesionTOADS scores. This difference was found to be 12.6, with a 95% confidence interval of (9.6, 15.8), p-value $< 10^{-12}$. The OASIS empirical threshold was ranked higher

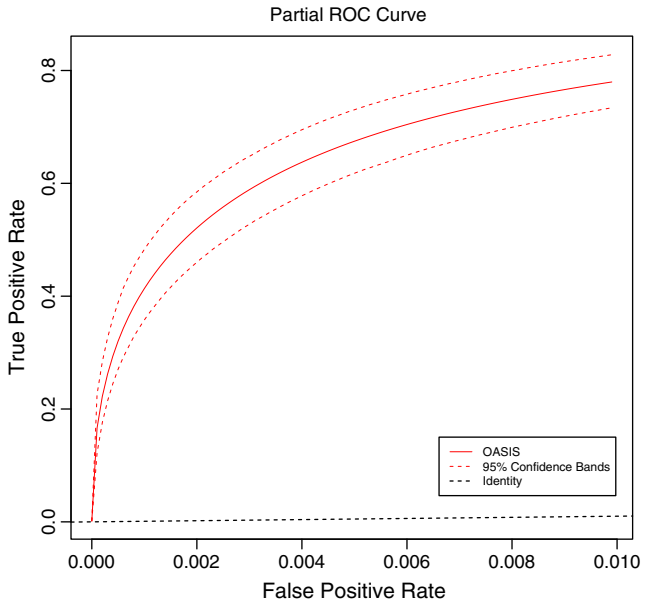

Fig. 4. Partial ROC curve for the voxel-level detection of lesions in the testing set of Validation Set 1 for different thresholds of the probability maps produced from OASIS for clinically relevant false positive rates of 1% and below. Bootstrapped 95% confidence bands are also provided. The vertical axis of the partial ROC curve shows the true positive rate (sensitivity) for a given threshold of the probability map and the horizontal axis shows the false positive rate (1 − specificity) for this threshold.

Table 2
Binary segmentation thresholds with false positive rate, sensitivity and DSC for Validation Set 1.

False positive rate	Sensitivity	Threshold value	DSC
1%	80%	0.10	0.55
0.75%	76%	0.12	0.58
0.5%	69%	0.16	0.61
0.25%	58%	0.23	0.59

than LesionTOADS segmentation in 73 (95% CI: [64, 81]) of the 98 studies or 74%(95% CI: [65%, 82%]). We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the rankings.

3.3. Validation Set 2: neuroradiologist rating results

Table 4 contains summary statistics for the scores from the neuroradiologist ratings of the three segmentations for all 189 studies. The OASIS Validation Set 1 threshold segmentations and the LesionTOADS segmentations have a much lower first quantile than the OASIS empirical threshold segmentations. For this analysis we focus mainly on the difference between the OASIS empirical threshold and the LesionTOADS segmentation, as the OASIS Validation Set 1 threshold did not perform well on this new data set. This was expected, as the probability map threshold needs to be adjusted to maintain the same false positive rate when moving to a new data set. We performed a paired *t*-test to assess the difference in the means of the OASIS empirical threshold scores and the LesionTOADS scores. This difference was found to be 16.6, with a 95% confidence interval of (13.3, 20.0), p-value $< 10^{-14}$. The OASIS empirical threshold was ranked higher than LesionTOADS segmentation in 146 (95% CI:[135, 157]) of the 189 cases or 77% (95% CI: [71%, 83%]). We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the rankings.

To assess rater reliability among the 20 duplicated MRI studies, we calculated the intraclass correlation coefficient: 0.61 (95% CI: [0.69, 0.81]). The rankings for the LesionTOADS images and the OASIS empirical threshold were preserved in the duplicate rankings for 17 of the 20 images (95% CI: [14, 20]). We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for both the intraclass correlation coefficients and the rankings.

3.4. Validation Set 2: rater agreement with neuroradiologist, neurologist, and radiologist

Table 5 contains summary statistics for the scores from the neuroradiologist, neurologist, and radiologist ratings of the three segmentations for the set of 50 studies selected to asses rater reliability. Fig. 5 shows a notched box plot for each rater of these findings. From the box plot we see that there is a statistically significant difference between the medians for all three segmentations for the neuroradiologist and neurologist. There was not a statistically significant difference in the medians of the scores for the three segmentations by the radiologist. Moreover, all three raters indicated that the OASIS Validation Set 1 segmentations and the LesionTOADS segmentations have a much lower first quantile than the OASIS empirical threshold segmentations. The outliers in the boxplots can be explained as either errors in processing, such as registration or bad artifacts, or as studies that none of the segmentation methods performed well on. We did not remove these studies from the analysis, because we want to assess the performance of OASIS in the setting of an image processing pipeline, where images may not be properly registered or may contain artifacts.

Again, we will focus mainly on the difference between the OASIS empirical threshold and the LesionTOADS segmentation. We performed a paired *t*-test to assess the difference in the means of the OASIS empirical threshold scores and the LesionTOADS scores. These differences can be found in Table 5. The mean for the OASIS empirical threshold was greater than the mean for the LesionTOADS scores for all three raters. This difference was found to be statistically significant for both the neuroradiologist and neurologist, (p-values $< 10^{-4}$ and $< 10^{-3}$, respectively), but not for the radiologist, (p-value 0.5). The neuroradiologist and the neurologist tended to spread their scores more, and this allowed better comparison of the segmentation algorithms. Table 5 also shows the percentage of time the OASIS empirical threshold was ranked higher than LesionTOADS segmentation in the 50 studies. We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the rankings.

To assess rater reliability among the 5 duplicated MRI studies, we calculated the intraclass correlation coefficient and the number of times the rankings for the LesionTOADS images and the OASIS empirical threshold were preserved. We nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for both the intraclass correlation coefficients and the rankings. For the neuroradiologist, the intraclass correlation coefficient for the 5 repeated studies is 0.55 (0.21, 0.82) and the number of preserved rankings is 4 (2,5). For the neurologist, 0.32 (−0.10, 0.68) and 4 (2,5). For the radiologist, −0.38 (−0.35, 0.71) and 2 (0,4). The repeated rankings for each rater for the 5 subjects are reported in the Appendix.

We calculated the rater agreement for the ranking of the OASIS empirical threshold versus LesionTOADS. We decided to use the rankings of the scores to assess rater agreement rather than the scores themselves, because, as shown from the intraclass correlation coefficient, the scores are not very reliable, while the order in which the observers rank the segmentations, on the other hand, is quite reliable. We calculated the kappa statistic to assess the reliability of the rankings for each pair of raters and nonparametrically bootstrapped with replacement the subjects to produce the confidence intervals for the kappa statistics. The kappa statistic for the rater agreement between the neuroradiologist and the neurologist was 0.47 (0.20, 0.75), the neuroradiologist and radiologist 0.02 (−0.26, 0.30) and the neurologist and radiologist −0.09 (−0.37, 0.19).

Table 3
Volume of false positive lesion in healthy volunteers and MS subjects from Validation Set 1 (in cm^3); the actual mean lesion volume is 0 cm^3 for healthy volunteers and 11.2 cm^3 (IQR: [1.7 cm^3, 16.6 cm^3]) for MS subjects.

Threshold value	Healthy mean (IQR)	MS mean (IQR)
0.10	8.6 (4.6, 10.6)	10.9 (7.6, 13.6)
0.12	6.7 (3.1, 8.2)	8.0 (5.2, 10.3)
0.16	4.3 (1.5, 5.7)	5.2 (3.0, 7.0)
0.23	2.2 (.7, 2.8)	2.5 (1.2, 3.5)

Table 4
Summary statistics of image ratings of Validation Set 2 for neuroradiologist on 189 studies.

	OASIS	OASIS	LesionTOADS
	Validation Set 1 threshold	Empirical threshold	
Minimum	3.7	3.7	2.7
1st quantile	27.3	55.7	21.7
Median	42.0	68.3	51.0
Mean	43.2	64.1	47.5
3rd quantile	57.7	76.3	71.0
Maximum	99.3	99.0	97.3

4. Discussion

OASIS may be used to assist or even replace manual segmentation of MS lesions in the brain. After training and adjustment of the population level threshold, our fully automatic method does not require

Table 5
Mean and standard deviation of the rating from the neuroradiologist, neurologist, and radiologist for OASIS Validation Set 1 threshold, OASIS empirical threshold and LesionTOADS on 50 studies from Validation Set 2; mean difference between OASIS empirical threshold and LesionTOADS and percentage of times OASIS was ranked higher than LesionTOADS on these images.

	OASIS Validation Set 1 Mean (SD)	OASIS Empirical Mean (SD)	LesionTOADS Mean (SD)	Mean Difference (95% CI)	Percentage Rank (95% CI)
Neuroradiologist	46.3 (22.0)	66.1 (20.2)	47.3 (27.2)	18.7 (11.2, 26.3)	76% (64%, 88%)
Neurologist	48.7 (24.3)	73.1 (18.5)	56.6 (26.0)	16.5 (7.0, 25.9)	66% (52%, 78%)
Radiologist	71.6 (19.6)	74.1 (17.9)	71.8 (16.5)	2.3 (−4.2, 8.8)	52% (38%, 66%)

human input and avoids the variability introduced by manual segmentation. Using the explicit form of the statistical model, OASIS can easily be adapted and trained for cases where more or fewer imaging sequences are available.

With the OASIS model, a recalibration of the population-level segmentation threshold is necessary for each new data set but can be done on a fairly limited number of subjects, as in the example from this paper. A recalibration of the population-level segmentation threshold is necessary for each new data set but can be done on a fairly limited number of subjects, as in the example from this paper. A set of subjects is required to tune this population level threshold, therefore fully automatic segmentation of a single study from a new imaging center may not be feasible with the OASIS model. However, in these cases the threshold can be adjusted very quickly manually (2–5 min) by visual inspection of 3–4 slices by adjusting just one parameter. When using an ROC curve for classification, thresholds for subpopulations with different covariate values may need to be defined differently in order to keep false positive rates the same across those subpopulations (Pepe, 2003). Therefore, it was expected that the ROC threshold would need to be adjusted to maintain the same false positive rate from Validation Set 1 in Validation Set 2. This threshold is the only tuning parameter in OASIS that must be adjusted when moving to a new data set, and this adjustment is very fast and intuitive to make and does not require multiple iterations of segmentations. We believe that OASIS holds promise for use in multicenter MRI studies, with adjustment of the population level threshold for each site.

Future work includes further validation of OASIS under changes in imaging center and protocol and to also show the reproducibility of the OASIS segmentations. One resource for this is the MS Lesion Segmentation Challenge (Styner et al., 2008), a common database for MS lesion segmentation algorithms. We plan to do further validation with this database as well as with volumes from additional imaging centers. For this analysis we did not have scan-rescan MRI available. These are crucial for assessing the reproducibility of the method, and we plan to acquire these in the future.

In contrast to many automatic segmentation techniques, OASIS is computationally fast. While training the model on the 131 studies from Validation Set 1 takes five hours on a standard workstation, this process is only conducted once. The results from this are summarized as the two sets of 21 coefficients in model [1]. Also, the model may be trained on fewer studies, as shown in the partial ROC analysis within Validation Set 1; the performance of the model remains stable when trained on subsets of 20 studies. Using this fitted model to generate a probability map of the entire brain from a set of new images takes only 30 min. These times are for standard workstations and are expected to drop dramatically with multi-core parallel computing and improved technologies. The Gaussian smoothing is the slowest step of the algorithm, and these computations can be parallelized to substantially decrease the time of the entire algorithm to approximately 5 min.

After making the image ratings for Validation Set 2, the neuroradiologist was unblinded and reviewed the three segmentations, providing comments about the strengths and weaknesses of each. The OASIS empirical threshold performed much better than the OASIS Validation Set 1 threshold. The neuroradiologist reported a preference for the smoothness of the OASIS segmentations in contrast to the LesionTOADS segmentation, which often appeared speckled. The OASIS segmentations often had artifacts in the pineal glands and the choroid plexus of the ventricles. This may be explained by the fact that OASIS was trained on FLAIR images acquired before a gadolinium-based contrast agent was administered to the patient, while the validation was done with FLAIR images that were acquired after gadolinium administration. Voxels in the choroid plexus and pineal glands, which enhance with gadolinium, were brighter and were thus misclassified as lesion. LesionTOADS does not make a similar error, as it imposes topological constraints that preclude these structures from being identified as lesions. Further refinements of OASIS may account for such complex

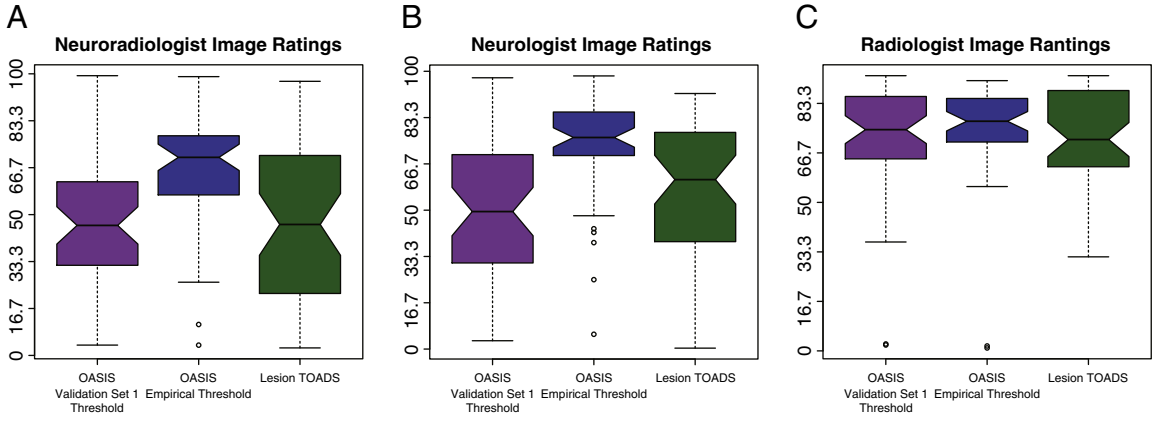

Fig. 5. Notched box plot of the results from the neuroradiologist, neurologist, and radiologist image ratings for segmentations of the 50 MRI studies from Validation Set 2: the OASIS Validation Set 1 threshold segmentations, the OASIS empirically adjusted threshold segmentations, and the LesionTOADS segmentations.

changes of protocol. The LesionTOADS segmentations were more variable than those of OASIS and did not perform well on cases with low lesion load. The OASIS segmentation had systematic errors in the medial frontal cortex and the brainstem. On the other hand, LesionTOADS avoided false positives in the brainstem because it only segments lesions in the cerebrum. Fig. 6 shows a slice from a subject with an example of a lesion that OASIS segments in the cerebellum. Fig. 6A shows a single slice of the FLAIR volume, Fig. 6B shows a single slice of the T1-weighted volume, Fig. 6C shows the LesionTOADS segmentation of the slice, and Fig. 6D shows the OASIS segmentation of the slice. LesionTOADS does not segment the cerebellum, whereas OASIS does not restrict the areas that it segments and is able to find the lesion in this slice.

OASIS is not an atlas-based method and therefore does not take into account anatomical information during segmentation, such as tissue class. Further incorporation of anatomical information, such as the tissue class segmentations from LesionTOADS, may help to avoid lesions false positives in areas where we have prior knowledge that lesion presence is low and where OASIS made systematic false positives, such as the medial frontal cortex and the brainstem. Also, this could be used to help with the false positives in the pineal glands and the choroid plexus of the ventricles in the post-contrast FLAIR as these are areas where lesions do not occur in MS.

The smoothed images used in OASIS are similar to the use of smoothed images for inhomogeneity correction in MRI. For inhomogeneity correction, an image is smoothed to suppress the details of the image and then the original image is divided by this smoothed image in order to correct the image inhomogeneity (Axel et al., 1987). Our method differs from this in that we do not divide the original image by the smoothed volume. Instead we use the smoothed volume as a covariate in our model. We also use multiresolution smoothed volumes, in contrast to just one smoothed volume for correction.

Other methods of capturing inhomogeneities may be used in the OASIS model as an alternative to the smoothed volumes. Alternative smoothers may be used instead of the Gaussian kernel and may be more appropriate in other applications. We decided to use the Gaussian filter because it is widely used, can be applied to any image, and is relatively computationally fast. The OASIS modeling framework is very flexible, however, and can be adapted for other methods of capturing the bias field and regional intensity variation.

We used the 15th percentile of FLAIR intensities in the brain to create the brain tissue mask. Other segmentations can be used to

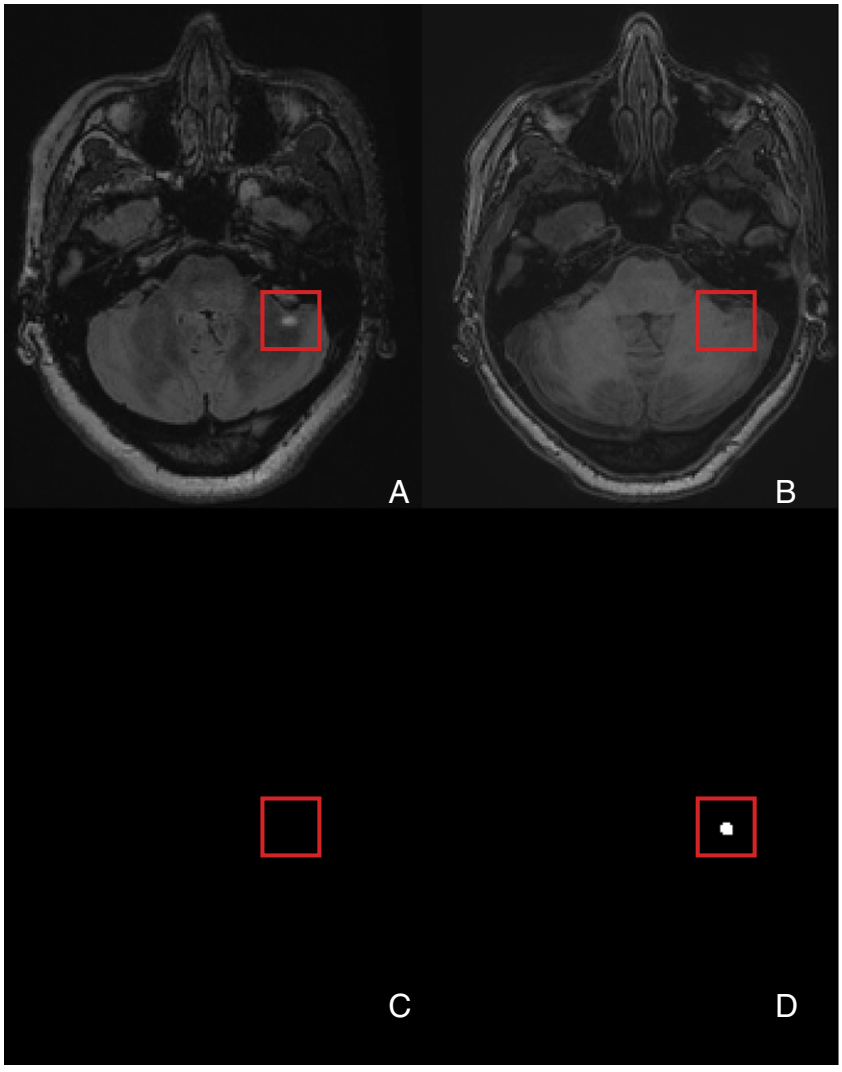

Fig. 6. Example of a cerebellum lesion classified using OASIS in Validation Set 2: A. FLAIR volume; B. T1-weighted volume; C. LesionTOADS segmentation; and D. OASIS empirically adjusted threshold segmentation.

remove the CSF. We used the 15th percentile of FLAIR intensities because it is fast and performed well in this application.

Lesions that are hypointense on FLAIR, because of high free water content, are not detected by OASIS. The method models only candidate voxels, the top 15% of voxels in the cerebral matter-masked FLAIR volume, to minimize the number of false positives. In the FLAIR volume, such lesions are characterized by hypointensities in the center of a lesion and hyperintensities around the edges. Therefore the center of the lesions is excluded from the candidate voxels. Future work includes expanding the OASIS model to segment these lesions. This could be done by fitting another OASIS model trained only on lesion voxels that appear hypointense in FLAIR lesions. The binary segmentations from the original OASIS model and this model could then be combined to produce a complete lesion segmentation.

Like other voxel-based methods, OASIS is sensitive to major misregistrations within an MRI study. However, in part because it incorporates spatial smoothing, OASIS is not sensitive to minor errors in registration. By simultaneously comparing data from multiple sequences and only considering candidate voxels, OASIS is able to distinguish between artifacts and lesion.

OASIS uses a voxel-level model for assessing the outcome. The assumption of independence between voxels is imperfect, as lesions consist of clusters of voxels. In this work we use smoothing in the smoothed volumes and smoothing of the predicted probabilities of the model to incorporate the spatial nature of the data. Nevertheless, further incorporation of neighboring voxel information is warranted.

Acknowledgments

The authors would like to thank Colin Shea, the Neuroimmunology Branch clinical group and the technicians at the NIH and Kirby Center who were instrumental in helping to collect and process the study data. This research was partially supported by the Intramural Research Program of NINDS, NINDS R01 NS070906, NINDS R01 NS060910, and NIBIB R01 EB012547.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2013.03.002.

References

Adler, D., Glaeser, C., Nenadic, O., OehlschlŠgel, J., Zucchini, W., 2011. ff: memory-efficient storage of large data on disk and fast access functions. R Package version 2.2–2. http://CRAN.R-project.org/package=ff.

Alfano, B., Brunetti, A., Larobina, M., Quarantelli, M., Tedeschi, E., Ciarmiello, A., Covelli, E.M., Salvatore, M., 2000. Automated segmentation and measurement of global white matter lesion volume in patients with multiple sclerosis. Journal of Magnetic Resonance Imaging 12 (6), 799–807.

Anbeek, P., Vincken, K.L., van Osch, M.J., Bisschops, R.H., van der Grond, J., 2004. Probabilistic segmentation of white matter lesions in MRimaging. NeuroImage 21 (3), 1037–1044.

Anbeek, P., Vincken, K.L., van Bochove, G.S., van Osch, M.J., van der Grond, J., 2005. Probabilistic segmentation of brain tissue inMR imaging. NeuroImage 27 (4), 795–804.

Anbeek, P., Vincken, K., Viergever, M., 2008. Automated MS-lesion segmentation by k-nearest neighbor classification. Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge, pp. 1–8.

Axel, L., Constantini, J., Listerud, J., 1987. Technical note: intensity correction in surface-coil MR imaging. AJR 148, 418–420.

Bordier, C., Dojat, M., Lafaye De Micheaux, P., 2009. AnalyzeFMRI: an R package to perform statistical analysis on FMRI data sets. R Package, version 1.1-12. http://CRAN.R-project.org/package=AnalyzeFMRI.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M., 1999. Global, voxel and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Transactions on Medical Imaging 18 (1), 32 Ð 41.

Carass, A., Cuzzocreo, J., Wheeler, M.B., Bazin, P.L., Resnick, S.M., Prince, J.L., 2011. Simple paradigm for extra-cerebral tissue removal: algorithm and analysis. NeuroImage 56 (4), 1982–1992.

Datta, S., Sajja, B.R., He, R., Wolinsky, J.S., Gupta, R.K., Narayana, P.A., 2006. Segmentation and quantification of black holes in multiple sclerosis. NeuroImage 29 (2), 467–474.

Dice, L., 1945. Measures of the amount of ecologic association between species. Ecology 25 (3), 297–302.

Dinh, T.A., Silander, T., Tchoyoson Lim, C.C., Leong, T.Y., 2012. An automated pathological class level annotation system for volumetric brain images. AMIA Annual Symposium Proceedings 1202–1210.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Medical Image Analysis 17 (1), 1–18.

Goldberg-Zimring, D., Achiron, A., Miron, S., Faibel, M., Azhari, H., 1998. Automated detection and characterization of multiple sclerosis lesions in brain MR images. Journal of Magnetic Resonance Imaging 16 (3), 311–318.

Hou, Z., 2006. A review on MR image intensity inhomogeneity correction. International Journal of Biomedical Imaging (1), 1–11.

Karimaghaloo, Z., Shah, M., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T., 2012. Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields. IEEE Transactions on Medical Imaging 31 (6), 1181–1194.

Lecoeur, J., FerrŽ, J.C., Barillot, C., 2009. Optimized supervised segmentation of MS lesions from multispectral MRIs. Work. Med. Image Anal. Mult. Scler, pp. 5–14.

Lee, C., Schmidt, M., Murtha, A., Bistritz, A., Sander, J., Greiner, R., 2005. Segmenting brain tumor with conditional random fields and support vector machines. Proceedings of Workshop on Computer Vision for Biomedical Image Applications atInternational Conference on Computer Vision.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrentá, L., Rovira, A., 2011. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. Information Sciences 186 (1), 164–185.

Lucas, B.C., Bogovic, J.A., Carass, A., Bazin, P.L., Prince, J.L., Pham, D.L., Landman, B.A., 2010. The Java Image Science Toolkit (JIST) for rapid prototyping and publishing of neuroimaging software. Neuroinformatics 8 (1), 5–17.

Lumley, T., 2009. biglm: bounded memory linear and generalized linear models. R package version 0.7. http://CRAN.R-project.org/package=biglm.

McAuliffe, M., Lalonde, F., McGarry, D., Gandler, W., Csaky, K., Trus, B., 2001. Medical image processing, analysis and visualization in clinical research. Proceedings of the 14th IEEE Symposium on Computer-Based Medical Systems (CBMS 2001), pp. 381–386.

Morra, J., Tu, Z., Toga, A., Thompson, P., 2008. Automatic segmentation of MS lesions using a contextual model for the MICCAI grand challenge. Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge, pp. 1–7.

Obuchowski, N.A., 2003. Receiver operating characteristic curves and their use in radiology. Radiology 299 (1), 3–8.

Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.

Rovira, A., León, A., 2008. MR in the diagnosis and monitoring of multiple sclerosis: an overview. European Journal of Radiology 67 (3), 409–414.

Rovira, A., Swanton, J., Tintor, M., Huerga, E., Barkhof, F., Filippi, M., Frederiksen, J.L., Langkilde, A., Miszkiel, K., Polman, C., Rovaris, M., Sastre-Garriga, J., Miller, D., Montalban, X., 2009. A single, early magnetic resonance imaging study in the diagnosis of multiple sclerosis. Archives of Neurology 66 (5), 587–592.

Sahraian, A.M., Radue, E.W., 2007. MRI Atlas of MS Lesions. Springer 178.

Sajja, B.R., Datta, S., He, R., Mehta, M., Gupta, R.K., Wolinsky, J.S., Narayana, P.A., 2006. Unified approach for multiple sclerosis lesion segmentation on brain MRI. Annals of Biomedical Engineering 34 (1), 142–151.

Scully, M., Magnotta, V., Gasparovic, C., Pelligrimo, P., Feis, D., Bockholt, H., 2008. 3D segmentation in the clinic: a grand challenge II at MICCAI 2008 — MS lesion segmentation. Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge, pp. 1–9.

Shiee, N., Bazin, P., Cuzzocreo, J.L., Reich, D.S., Calabresi, P.A., Pham, D.L., 2008a. Topologically constrained segmentation of brain images with multiple sclerosis lesions. Work. Med. Image Anal. Mult. Scler, pp. 71–81.

Shiee, N., Bazin, P., Pham, D.L., 2008b. Multiple sclerosis lesion segmentation using statistical and topological atlases. Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge, pp. 1–10.

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

Shinohara, R.T., Crainiceanu, C.M., Caffo, B.S., Gaitan, M.I., Reich, D.S., 2011. Population-wide principal component-based quantification of blood–brain-barrier dynamics in multiple sclerosis. NeuroImage 57 (4), 1430–1446.

Shinohara, R.T., Goldsmith, J., Mateen, F.J., Crainiceanu, C., Reich, D.S., 2012. Predicting breakdown of the blood–brain barrier in multiple sclerosis without contrast agents. AJNR. American Journal of Neuroradiology 33 (8), 1586–1590.

Simmons, A., Tofts, P.S., Barker, G.J., Arridge, S.R., 1994. Sources of intensity nonuniformity in spin echo images. Magnetic Resonance in Medicine (32), 121–128.

Simon, J.H., Li, D., Traboulsee, A., Coyle, P.K., Arnold, D.L., Barkhof, F., Frank, J.A., Grossman, R., Paty, D.W., Radue, E.W., Wolinsky, J.S., 2006. Standardized MR imaging protocol for multiple sclerosis: consortium of MS Centers consensus guidelines. American Journal of Neuroradiology 27 (2), 455–461.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2009. ROCR: Visualizing the performance of scoring classifiers. R package version 1.0–4. http://CRAN.R-project.org/package=ROCR.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17 (1), 87–97.

Sormani, M.P., Bonzano, L., Roccatagliata, L., Cutter, G.R., Mancardi, G.L., Bruzzi, P., 2009. Magnetic resonance imaging as a potential surrogate for relapse in multiple sclerosis: a meta-analytic approach. Annals of Neurology 65 (3), 270–277.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Jewells, V., Warfield, S., 2008. Editorial: 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. Grand Challenge Work.: Mult. Scler. Lesion Segm. Challenge, pp. 1–8.

Subbanna, N., Shah, M., Francis, S.J., Narayannan, S., Collins, D.L., Arnold, D.L., Arbel, T., 2009. MS lesion segmentation using Markov Random Fields. Work. Med. Image Anal. Mult. Scler. , pp. 15–26.

Sweeney, E.M., Shinohara, R.T., Shea, C.D., Reich, D.S., Crainiceanu, C.M., 2013. Automatic lesion incidence estimation and detection using multisequence longitudinal MRIs. AJNR. American Journal of Neuroradiology 34 (1), 68–73.