Contents lists available at SciVerse ScienceDirect

NeuroImage: Clinical

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

Improving the clinical correlation of multiple sclerosis black hole volume change by paired-scan analysis☆

Roger C. Tam [a,b,*], Anthony Traboulsee [b,1], Andrew Riddehough [b,1], David K.B. Li [a,b,1]

[a] Department of Radiology, University of British Columbia, Vancouver, Canada
[b] Division of Neurology, University of British Columbia, Vancouver, Canada

ARTICLE INFO

Article history:
Received 13 July 2012
Received in revised form 23 August 2012
Accepted 27 August 2012
Available online 20 September 2012

Keywords:
Multiple sclerosis
Black holes
Image segmentation
Longitudinal analysis
Volume change
Clinical correlation

ABSTRACT

The change in T_1-hypointense lesion ("black hole") volume is an important marker of pathological progression in multiple sclerosis (MS). Black hole boundaries often have low contrast and are difficult to determine accurately and most (semi-)automated segmentation methods first compute the T_2-hyperintense lesions, which are a superset of the black holes and are typically more distinct, to form a search space for the T_1w lesions. Two main potential sources of measurement noise in longitudinal black hole volume computation are partial volume and variability in the T_2w lesion segmentation. A paired analysis approach is proposed herein that uses registration to equalize partial volume and lesion mask processing to combine T_2w lesion segmentations across time. The scans of 247 MS patients are used to compare a selected black hole computation method with an enhanced version incorporating paired analysis, using rank correlation to a clinical variable (MS functional composite) as the primary outcome measure. The comparison is done at nine different levels of intensity as a previous study suggests that darker black holes may yield stronger correlations. The results demonstrate that paired analysis can strongly improve longitudinal correlation (from -0.148 to -0.303 in this sample) and may produce segmentations that are more sensitive to clinically relevant changes.

1. Introduction

"Black holes" (BHs) in multiple sclerosis (MS) are typically defined as white matter lesions with a hypointense appearance relative to normal-appearing white matter (NAWM) on a T_1-weighted (T_1w) MRI and that also appear hyperintense in the corresponding T_2-weighted (T_2w) image. The general importance of T_1-hypointensity in MS pathology is acknowledged and there is a substantial body of histopathological evidence that supports chronic BHs as being indicative of irreversible demyelination and axonal damage (Neema et al., 2007; Sahraian et al., 2010; van den Elskamp et al., 2008). As a result, BH evolution is considered one of the most promising imaging endpoints in MS clinical trials (Barkhof et al., 2009). However, the relationship between BH measures, the most common being T_1w lesion volume, and clinical features remains unclear (Naismith and Cross, 2005). Previous studies investigating the correlation between BH volume and MS disability measures have produced inconsistent results (Naismith and Cross, 2005; Sahraian et al., 2010), which can be partly attributed to the differences in the image analysis methods used (Neema et al., 2007).

BHs are generally difficult to identify and measure because their boundaries are often of low contrast. They are typically less distinct than T_2w lesions (Horsfield et al., 2007; Zhao et al., 2000), for which segmentation techniques are still an active area of research. Part of the difficulty lies in the fact that BHs are often inhomogeneous, and the intensity variations that can appear within and between BHs make consistent identification and delineation challenging. Longitudinally, BH intensity typically changes as the lesion evolves, becoming darker with greater injury, while reduced edema or remyelination can cause the signal to increase (Sahraian et al., 2010). While the degree of T_1-hypointensity is useful in that it can reflect the amount of tissue destruction (Barkhof et al., 2003; Riva et al., 2009; van Walderveen et al., 1998), such variations can also be artifactual, with partial volume being a large contributor, especially when the clinical standard slice thickness of 3 mm is used. Due to the fact that BHs are generally more difficult to delineate, and because they are always a subset of the regions occupied by T_2w lesions, most automated methods for computing BH volume use the T_2w lesions to help define a search space for the hypointensities on the T_1w scan (e.g., Datta et al., 2006; Wu et al., 2006). A natural consequence is that variability in the T_2w lesion segmentation can directly impact the measurement of BH volume. Even if the T_2w lesion segmentation

* Corresponding author at: MS/MRI Research Group, 203-2386 East Mall, Vancouver, British Columbia, Canada V6T 1Z3. Tel.: +1 604 827 5381; fax: +1 604 822 7877.
E-mail addresses: Roger.Tam@ubc.ca (R.C. Tam), T.Traboulsee@ubc.ca (A. Traboulsee), Andrew@msmri.medicine.ubc.ca (A. Riddehough), David.Li@ubc.ca (D.K.B. Li).
[1] MS/MRI Research Group, 203-2386 East Mall, Vancouver, British Columbia, Canada V6T 1Z3.

2213-1582/$ – see front matter © 2012 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.nicl.2012.08.004

is accurate for each individual scan, the variability between scans may negatively impact the measurement of T_1w lesion volume change.

To measure the change in BH volume between a baseline scan and follow-up scan, the variability induced by partial volume and T_2w segmentation can be potentially reduced by performing a paired analysis that uses both scans together. In this study, we use a large set of MRIs of MS patients to investigate the impact of: 1) image registration to equalize the partial volume across scans and 2) combining the T_2w lesion masks from the baseline and follow-up scans, in a process we term *mask averaging*, to produce a unified search space for the BHs. As an outcome measure, we use the rank correlation between the change in global T_1w lesion volume and MS disability status as quantified by the MS functional composite (MSFC) (Fischer et al., 1999) to determine if the proposed methodology has utility in a clinical context. Our hypothesis is that using registration and mask averaging may allow a given BH segmentation algorithm to be more sensitive to the real change induced by pathological progression, thereby resulting in stronger longitudinal clinical correlations. We perform comparisons between the unpaired and paired methods at nine different intensity thresholds. The motivation for analyzing BHs at different intensities comes from a recent small study (Tam et al., 2011) in which we observed that the cross-sectional correlation between BH volume and clinical disability can be strongly influenced by the intensity range used to define the BHs, and limiting the measurement to the darker regions can yield stronger correlations. In this study, we incorporate a similar analysis, but with a much larger data set and an added longitudinal dimension.

2. Materials and methods

2.1. Data

The MRIs of 247 patients with secondary progressive MS (SPMS) participating in a clinical trial were used. The data set contains two visits per patient, acquired approximately two years apart. The scans were obtained at 14 scanning sites and each visit set includes T_1w, T_2w and proton density-weighted (PDw) scans. Contrast-enhanced T_1w scans were used in the BH volume calculations in order to avoid including the enhanced regions which are indicative of active inflammation. The T_1w scans were acquired with a repetition time of 600.0–800.0 ms and echo time of 9.0–20.0 ms. The T_2w and PDw scans were acquired in a dual-echo sequence with a repetition time of 2500.0–3000.0 ms, first echo time of 8.4–20.0 ms and second echo time of 60.6–98.0 ms. All of the images have $256 \times 256 \times 50$ voxels with the size of $0.937 \times 0.937 \times 3.0$ mm, and no interslice gap. Each patient was assessed by a qualified neurologist to produce an MSFC score, which measures an MS patient's physical and cognitive abilities relative to a population distribution, and is expressed as the number of positive or negative standard deviations from the mean. Table 1 shows the summary statistics for the baseline, two-year and change in MSFC scores for the patient sample. Appropriate ethics approval was obtained from the institutional ethics review boards for all data acquired in this study.

Table 1
Summary statistics for the multiple sclerosis functional composite (MSFC) scores and T_2w lesion volumes (mm^3) in the current sample of 247 patients. The mean, standard deviation (SD) and interquartile range (IQR) are given for the baseline, two-year and change values. The cross-sectional Spearman correlation between T_2w lesion volume and MSFC is −0.481 ($p=1.03 \times 10^{-15}$) at baseline and −0.459 ($p=2.72 \times 10^{-14}$) at follow-up. The longitudinal correlation is not statistically significant ($p=0.109$).

	Baseline	Two-year	Change
MSFC mean	0.118	−0.149	−0.266
MSFC SD	0.761	1.458	1.251
MSFC IQR	0.575	0.601	0.210
T_2w lesion mean	10947.30	12014.04	1066.75
T_2w lesion SD	9839.03	10506.89	2660.59
T_2w lesion IQR	7685.74	9122.54	1097.02

To evaluate the effect of registration and T_2w mask averaging, we analyzed our test data using three different methods:

1) Single-scan (unpaired): a method that segments the BHs on each T_1w scan individually. As explained below, the method uses the corresponding T_2w and PDw scans to compute a T_2w lesion mask which forms a search space for the T_1-hypointensities.
2) Paired using registration: a paired method that registers the baseline and follow-up T_1w scans for each patient, then applies the single-scan method, using an individual T_2w lesion mask for each scan.
3) Paired using registration + T_2w mask averaging: a paired method that registers the baseline and follow-up T_1w scans for each patient, combines the two T_2w lesion masks into one unified mask, then applies the single-scan method but using the combined mask for both time points.

2.2. Single-scan (unpaired) lesion segmentation

The BH segmentation process that can be applied to a single scan is detailed in previous work (Tam et al., 2011; McAusland et al., 2010) and diagrammed in Fig. 1, so it is only summarized here. The pipeline begins with the manual identification of the T_2w lesions via the placement of seed points, but the rest of the processing is fully automatic. We choose to use this level of interactivity because while T_2w/PDw MRIs are very sensitive to white matter abnormalities, they lack specificity and expert knowledge is required to distinguish MS lesions from other pathology (Filippi et al., 2005). After preprocessing the scans (Jones and Wong, 2002; Smith, 2002), two radiologists are asked to place one or more seed points to mark the location and approximate extent of each lesion visible on the T_2w and PDw scans. The radiologists follow a set of guidelines that is minimalistic and allows the seeding procedure to be efficient and intuitive. The seed points are processed by a customized Parzen windows (Parzen, 1962) classifier to estimate the intensity distribution of the lesions, and connected component and shape analyses are then used to compute the final T_2w lesion segmentations.

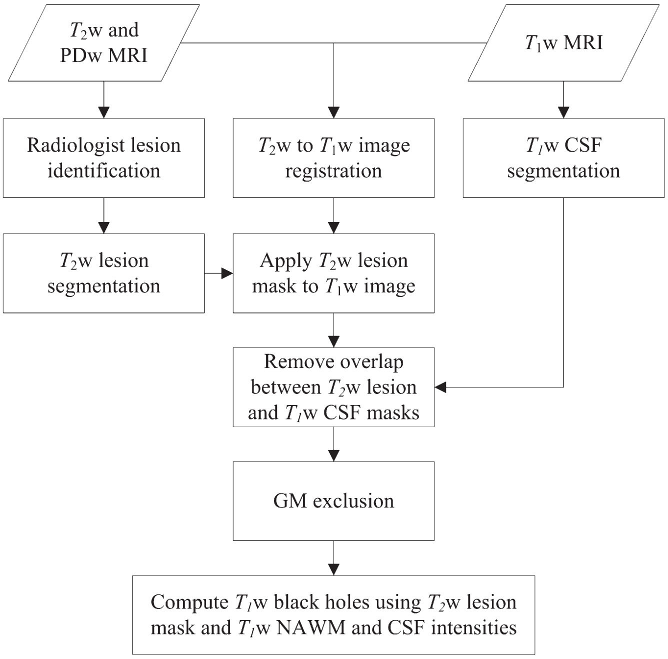
T2w and PDw MRI
T1w MRI
Radiologist lesion identification
T2w to T1w image registration
T1w CSF segmentation
T2w lesion segmentation
Apply T2w lesion mask to T1w image
Remove overlap between T2w lesion and T1w CSF masks
GM exclusion
Compute T1w black holes using T2w lesion mask and T1w NAWM and CSF intensities

Fig. 1. Overview of the black hole segmentation process. The only step that requires manual interaction is the identification of the T_2w lesions.

The algorithm uses the T_2w lesion masks to automatically extract the BHs from the T_1w scans. Each T_2w scan is first rigidly registered to the corresponding T_1w scan so that the binary T_2w lesion mask can be overlaid to define a search space for the T_1-hypointense regions. For each T_1w voxel that is not completely contained within the T_2w lesion mask, a partial volume value is determined by modeling the voxels as polyhedra and computing the volume of the intersection between the T_1w voxel and the closest transformed voxels of the T_2w lesion mask. The result is a value that characterizes the percentage of each T_1w voxel that is occupied by T_2w lesion tissue. This computation is particularly important because of the large anisotropy in the through-slice direction. Only the T_1w voxels with 25% or more of their volume covered by the T_2w mask are further considered for inclusion as a BH voxel. The T_2w mask intersection and thresholding procedures result in a largely accurate search space, but in some regions, the applied T_2w lesion mask can still intrude slightly (generally less than 3 pixels) into the cerebrospinal fluid (CSF) regions of the T_1w scan. To exclude CSF from contributing falsely to the BH volume, a two-step CSF classification procedure is applied to the T_1w image: 1) histogram-based thresholding to form a conservative CSF mask; 2) the CSF mask on each slice is refined using a 2D geodesic active contour (Caselles et al., 1997), which is a boundary model that deforms according to local gradient information, in this case expanding the CSF segmentation to push any T_2w lesion mask intrusions out of the CSF.

To extract the BHs that lie within the refined T_2w lesion mask at a given level of hypointensity, relative to the intensities of NAWM and CSF, which may vary from scan to scan, the following upper intensity threshold (t_x) is used to determine whether any given voxel x should be included:

$$t_x = l \times \left(i_{\mathrm{NAWM},x} - i_{\mathrm{CSF,slice}}\right) + i_{\mathrm{CSF,slice}}$$

where $i_{\mathrm{NAWM},x}$ is the NAWM intensity, taken as the mean of the T_1w intensities of the 20 WM voxels that are closest to x on the same slice and also outside of the T_2w lesion containing x, $i_{\mathrm{CSF,slice}}$ is the mean CSF intensity over the entire slice, and l is the key parameter that is varied to span the range between NAWM and CSF. This method of computing an individual threshold for every voxel is designed to maximize the use of local contrast information and also excludes the contrast-enhanced areas. Gray matter is distinguished from WM and excluded from the computation of $i_{\mathrm{NAWM},x}$ by applying a modified fuzzy clustering algorithm (McAusland et al., 2004) to the T_2w/PDw combined intensity space. Every voxel x that has intensity $i_x \leq t_x$ is counted as a BH voxel. For this study, we varied l from 0.90 (closest to NAWM) to 0.10 (closest to CSF) with a decrement of 0.10, resulting in BH volumes at nine different levels of maximum intensity. Fig. 2 shows examples of BHs segmented at three different intensity levels. From previous work (Tam et al., 2011), we determined that an l of 0.80, or just slightly hypointense, produces segmentations that most closely match the full visual extent of the BHs as manually traced by radiologists.

2.3. Paired lesion segmentation using registration

To equalize the partial volume between time points, we perform rigid registration between the baseline and follow-up images. Registration is done with the maximization of mutual information using Shannon entropy (Pluim et al., 2003) as the image similarity measure. The transformation parameters are computed by using the baseline image as the fixed image and the follow-up image as the moving image. To avoid asymmetric blurring during resampling, the resulting transformation is split into two halfway transformations, one "forward" and one "backward", that are applied to the baseline and follow-up images individually. Cubic spline interpolation (Meijering et al., 2001) is used for image resampling. The resulting images are aligned with the corresponding voxels in each image having undergone the same degree of interpolation. The single-scan BH segmentation method is then applied to both T_1w images using their individual T_2w masks.

2.4. Paired lesion segmentation using registration + T_2w lesion mask averaging

To remove variability in the T_2w lesion segmentation across time, we combine the T_2w lesion masks from the baseline and follow-up scans and apply the unified mask to both time points. First, rigid registration is performed as described above. For each voxel in the registered space, the baseline and follow-up images each have a T_2w mask value that represents the percentage of the voxel that is occupied by T_2w lesion. We compute a new T_2w lesion mask by using the mean of the two lesion mask values at each voxel. The single-scan BH segmentation method is then applied to the baseline and follow-up T_1w images, using the unified T_2w mask. The same lower mask value of 25% is used to threshold the unified mask as would be done for the individual masks. We could conceivably use the greater of the two mask values rather than the mean, but the mean produces a less aggressive mask that is less prone to false positives at the boundaries with CSF, and allows the same mask threshold and other parameter values to be used for a more direct comparison with unpaired processing. We also considered simply applying the follow-up masks to both time points, because lesion loads generally increase over time, so many of the baseline masks would be included in the follow-up masks. However, some lesions (especially T_2w) can resolve over time and some patients, especially those under medication, can experience in a reduction in lesion load. Therefore, we have chosen to use a combined mask to avoid biasing the segmentation toward either time point. Fig. 2 shows examples of T_2w masks and BHs produced with the three methods.

2.5. Statistical analysis

For each BH quantification method and each of the nine levels of maximum BH intensity, we computed the mean and coefficient of variation (CoV) of the T_1w lesion volumes at baseline, 2 years and of the changes in BH volume over 2 years. We chose to use the CoV, which is the ratio of the standard deviation (SD) over the mean, because both the mean and SD vary greatly across different intensity levels, and normalizing the SD by the mean facilitates comparison of the sample variation across different values of l. For each lesion quantification method and each intensity level, we also computed the cross-sectional and longitudinal Spearman correlations between BH volume and MSFC.

3. Results

3.1. T_2w lesion statistics

To give a general idea of the magnitude of the lesion changes, summary statistics of the T_2w lesion loads in this patient sample are given in Table 1. The cross-sectional Spearman correlation between T_2w lesion volume and the MSFC is −0.481 ($p = 1.03 \times 10^{-15}$) at baseline and −0.459 ($p = 2.72 \times 10^{-14}$) at follow-up. The longitudinal correlation is not statistically significant ($p = 0.109$).

3.2. Cross-sectional results

Table 2 shows the means, CoVs and cross-sectional MSFC correlations of the baseline BH volumes computed by the three methods at the nine different intensity levels. The differences between the three methods in the mean BH volumes computed at any intensity level are not statistically significant ($p > 0.05$), as evaluated by Wilcoxon

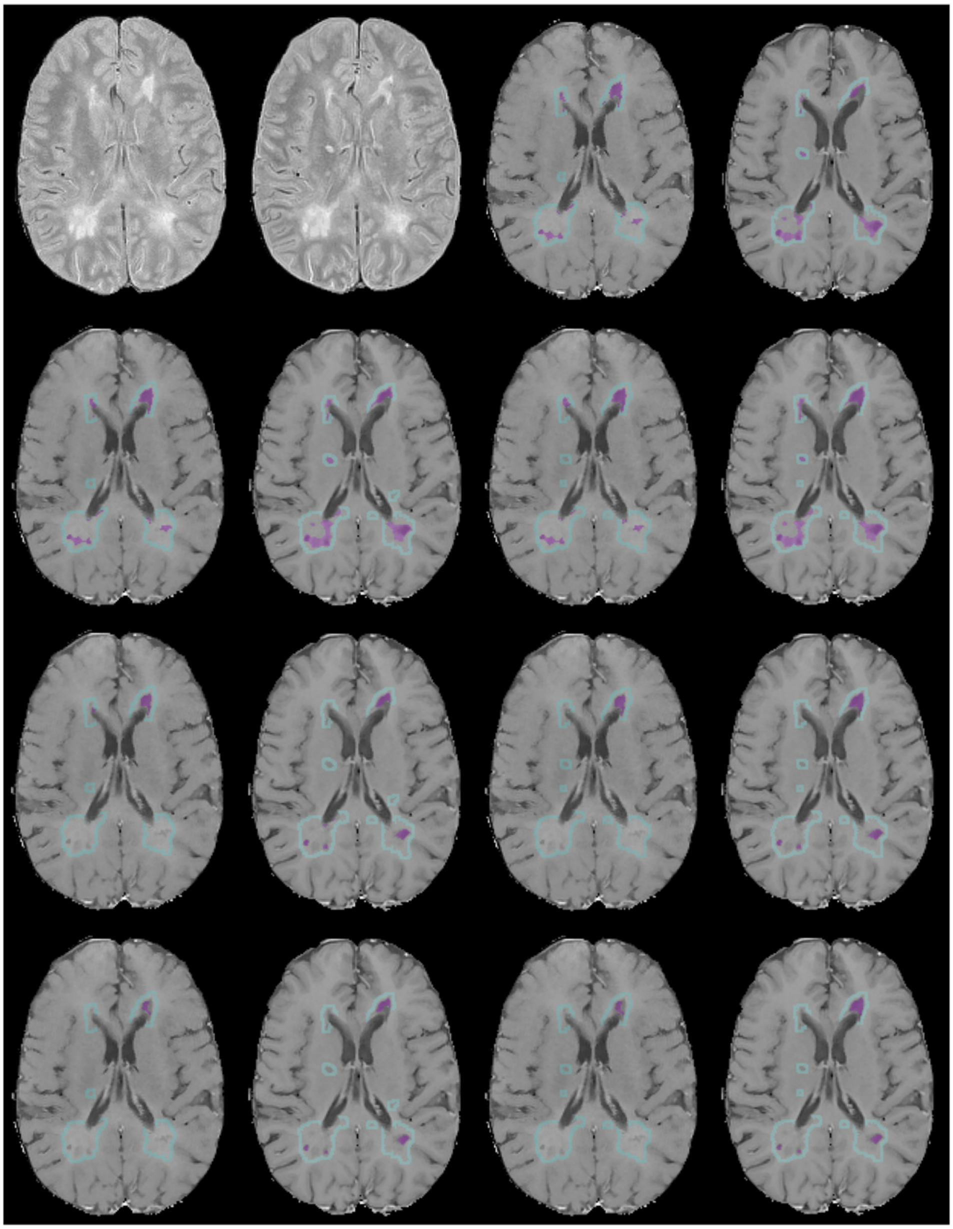

Fig. 2. Examples of T_2w lesion masks (cyan outlines) and resulting black holes (purple regions) produced with the unpaired, registration-only and registration + T_2w mask averaging methods. Top row, left two images: PDw images at baseline and 2 years, unregistered; top row, right two images: T_2w lesion masks and black holes computed at $l = 0.80$ with the unpaired (no registration) method on T_1w images at baseline and 2 years. Second row, left two images: T_2w lesion masks and black holes computed at $l = 0.80$ with registration-only on T_1w images at baseline and 2 years; second row, right two images: T_2w lesion masks and black holes computed at $l = 0.80$ with registration + T_2w mask averaging on T_1w images at baseline and 2 years. Third row, left two images: T_2w lesion masks and black holes computed at $l = 0.50$ with registration-only on T_1w images at baseline and 2 years; third row, right two images: T_2w lesion masks and black holes computed at $l = 0.50$ with registration + T_2w mask averaging on T_1w images at baseline and 2 years. Bottom row, left two images: T_2w lesion masks and black holes computed at $l = 0.30$ with registration-only on T_1w images at baseline and two years; bottom row, right two images: T_2w lesion masks and black holes computed at $l = 0.30$ with registration + T_2w mask averaging on T_1w images at baseline and 2 years.

rank-sum tests. For all three methods, the clinical correlations have a clear pattern of being the strongest when all hypointense areas are included ($l = 0.90$), and decreasing monotonically with decreasing BH intensity. All of the correlations are strongly statistically significant ($p < 0.0001$) and do not have obvious differences between methods.

Table 3 shows the means, CoVs and cross-sectional MSFC correlations of the two-year BH volumes computed by the three methods at the nine different intensity levels. The mean two-year T_1w lesion volumes are all larger than the corresponding baseline volumes, which is an expected finding in progressive MS patients. As with the baseline volumes, there are no statistically significant differences between the three methods at any intensity threshold. The cross-sectional clinical correlations have a similar pattern as the baseline results, with a monotonic decrease from a maximum at $l = 0.90$ and the same level of statistical significance ($p < 0.0001$), and do not have obvious differences between methods.

Despite the fact that the mean volumes computed by the three methods are not statistically different, the CoVs are large (>1), which would overwhelm any small effects, and it is still useful to examine the magnitudes and signs of the differences between methods to discover patterns that may help explain the longitudinal results. Table 4 shows the differences in mean volume between the three methods. The registration-only method computed BH volumes that are slightly larger (up to +1.8%) for $l = 0.90$ but lower for $l \leq 0.70$ (range: −1.5% to −9.6%, with generally larger percentage

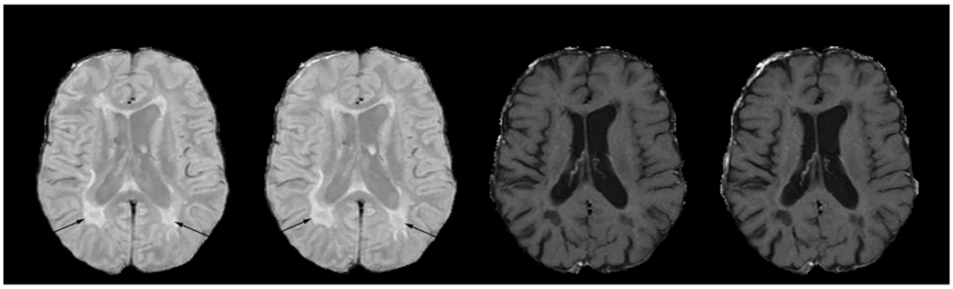

Fig. 3. PDw and T_1w images showing the evolution of MS lesions. From left to right: PDw image at baseline, PDw image at two years, T_1w image at baseline, T_1w image at two years. The arrows on the PDw images indicate two lesions with dark cores that become larger and darker over time. The intensity of the cores become closer to NAWM on the follow-up PDw image, while on T_1w the corresponding black holes continue to become larger and darker. This combination of intensity changes can cause underestimation of the black hole volume changes when the T_2w lesions are used as a search space.

difference for lower l) than the unpaired method, and affected the baseline and two-year volumes similarly. T_2w mask averaging slightly decreased the baseline volumes uniformly across all levels of l, with a range of −1.7% to −2.6%, compared to the registration-only method. Mask averaging also decreased the two-year volumes, compared to the registration-only method, for the higher levels of l (−6.6% and −3.5% for $l=0.90$ and 0.80, respectively), but increased the volumes for $l \leq 0.60$ at a progressively higher rate with decreasing l, from +2.1% at $l=0.60$ to +18.1% at $l=0.10$. Overall, the effect of registration + mask averaging produced lower baseline volumes compared to unpaired analysis and also produced lower two-year volumes for the higher values of l, but produced higher two-year volumes for the lower values of l.

3.3. Longitudinal results

Table 5 shows the means, CoVs and longitudinal MSFC correlations of the changes in BH volume over 2 years computed by the three methods at the nine different intensity levels. When evaluated by Wilcoxon rank-sum tests, two statistically significant differences between the three methods are observed in the volume changes computed. The registration + mask averaging method measured a lower mean change than the unpaired method (542.74 vs. 920.82 mm^3, $p=0.005$) and the registration-only method (542.74 vs. 850.52 mm^3, $p=0.002$), but only at $l=0.90$. Overall, the volume changes measured by the registration-only method are slightly lower than the unpaired method. The volume changes measured by registration + mask averaging are lower than the registration-only method for $l=0.90$ and 0.80, but are higher than both other methods for $l \leq 0.70$.

The key findings in this study are the longitudinal correlations between BH volume and MSFC, specifically as affected by the paired analysis methods. For the unpaired method, only the correlation at $l=0.30$ is statistically significant (−0.148, $p=0.020$). For the registration-only method, the number of intensity levels that yield significant correlations increases to four ($l=0.50$ to 0.20), with the strongest correlation at $l=0.40$ (−0.175, $p=0.006$). For the registration + mask averaging method, the correlations reach statistical significance for all values of l, and the magnitudes of the correlations are much higher overall, peaking at $l=0.30$ (−0.303, $p=1.27\times10^{-6}$) and decreasing monotonically in both value and significance on both sides of the maximum. Even the lowest

Table 2
Baseline T_1w lesion volume means (in mm^3), coefficients of variation (CoV), and cross-sectional rank correlations to MSFC for the unpaired, registration-only and registration + T_2w mask averaging methods. The variable l indicates the maximum intensity, relative to NAWM and CSF, used to define the black holes, with $l=0.90$ being closest to NAWM and therefore the most inclusive. A value of $l=0.80$ corresponds to the traditional black hole definition of including all visually hypointense voxels. The correlations are similar in value and significance for all three methods. For the correlation, *indicates $p<0.05$, **indicates $p<0.01$, ***indicates $p<0.001$ and ****indicates $p<0.0001$.

	Unpaired		Registration only		Registration + T_2w mask averaging	
l	Volume (CoV)	Spearman correlation	Volume (CoV)	Spearman correlation	Volume (CoV)	Spearman correlation
0.90	6092.42 (1.11)	-0.481^{****}	6201.10 (1.11)	-0.482^{****}	6045.66 (1.10)	-0.474^{****}
0.80	4173.54 (1.21)	-0.469^{****}	4177.04 (1.24)	-0.473^{****}	4078.53 (1.21)	-0.468^{****}
0.70	2785.74 (1.40)	-0.456^{****}	2743.43 (1.43)	-0.460^{****}	2695.27 (1.39)	-0.461^{****}
0.60	1944.84 (1.56)	-0.445^{****}	1889.18 (1.62)	-0.446^{****}	1856.61 (1.56)	-0.445^{****}
0.50	1353.96 (1.74)	-0.433^{****}	1294.37 (1.81)	-0.422^{****}	1271.31 (1.74)	-0.428^{****}
0.40	925.31 (1.93)	-0.405^{****}	869.73 (2.02)	-0.398^{****}	852.11 (1.94)	-0.416^{****}
0.30	603.53 (2.15)	-0.398^{****}	562.36 (2.25)	-0.383^{****}	547.68 (2.16)	-0.408^{****}
0.20	366.30 (2.42)	-0.358^{****}	338.09 (2.55)	-0.347^{****}	330.39 (2.45)	-0.374^{****}
0.10	204.74 (2.89)	-0.314^{****}	188.05 (3.04)	-0.320^{****}	184.78 (2.93)	-0.337^{****}

Table 3
Two-year T_1w lesion volume means (in mm^3), coefficients of variation (CoV), and cross-sectional rank correlations to MSFC for the unpaired, registration-only and registration + T_2w mask averaging methods. The variable l indicates the maximum intensity, relative to NAWM and CSF, used to define the black holes, with $l=0.90$ being closest to NAWM and therefore the most inclusive. A value of $l=0.80$ corresponds to the traditional black hole definition of including all visually hypointense voxels. The correlations are similar in value and significance for all three methods. For the correlation, *indicates $p<0.05$, **indicates $p<0.01$, ***indicates $p<0.001$ and ****indicates $p<0.0001$.

	Unpaired		Registration only		Registration + T_2w mask averaging	
l	Volume (CoV)	Spearman correlation	Volume (CoV)	Spearman correlation	Volume (CoV)	Spearman correlation
0.90	7013.25 (1.04)	-0.448^{****}	7051.63 (1.03)	-0.449^{****}	6588.40 (1.06)	-0.443^{****}
0.80	4797.05 (1.15)	-0.441^{****}	4761.70 (1.14)	-0.442^{****}	4597.25 (1.16)	-0.440^{****}
0.70	3200.95 (1.34)	-0.437^{****}	3139.40 (1.32)	-0.438^{****}	3125.76 (1.34)	-0.435^{****}
0.60	2237.43 (1.53)	-0.432^{****}	2159.03 (1.51)	-0.429^{****}	2204.31 (1.51)	-0.430^{****}
0.50	1569.96 (1.74)	-0.422^{****}	1486.85 (1.72)	-0.426^{****}	1558.23 (1.71)	-0.422^{****}
0.40	1084.70 (1.94)	-0.403^{****}	1010.40 (1.91)	-0.404^{****}	1092.05 (1.90)	-0.407^{****}
0.30	723.51 (2.11)	-0.374^{****}	657.87 (2.09)	-0.380^{****}	730.42 (2.10)	-0.391^{****}
0.20	456.68 (2.29)	-0.328^{****}	412.93 (2.26)	-0.318^{****}	469.24 (2.31)	-0.342^{****}
0.10	260.19 (2.61)	-0.281^{****}	238.90 (2.41)	-0.269^{****}	282.05 (2.55)	-0.298^{****}

Table 4
Differences in mean T_1w lesion volume in mm^3 between the three computation methods: unpaired (UP), registration-only (RO) and registration + T_2w mask averaging (RMA). The differences are not statistically significant ($p>0.05$) as evaluated by Wilcoxon rank-sum tests, but the patterns are useful for explaining the differences in the changes in volume over time as computed by the three methods (Table 5). The most notable pattern is that for the lower values of *l*, RMA computed lower baseline volumes but higher two-year volumes than the other methods, which increased the magnitude of the changes computed by RMA.

	Baseline volume differences in mm^3			Two-year volume differences in mm^3		
l	RO−UP	RMA−RO	RMA−UP	RO−UP	RMA−RO	RMA−UP
0.90	+108.7 (+1.8%)	−155.4 (−2.5%)	−46.8 (−0.8%)	+38.4 (+0.5%)	−463.2 (−6.6%)	−424.9 (−6.1%)
0.80	+3.5 (+0.1%)	−98.5 (−2.4%)	−95.0 (−2.3%)	−35.4 (−0.7%)	−164.4 (−3.5%)	−199.8 (−4.2%)
0.70	−42.3 (−1.5%)	−48.2 (−1.8%)	−90.5 (−3.2%)	−61.5 (−1.9%)	−13.6 (−0.4%)	−75.2 (−2.3%)
0.60	−55.7 (−2.9%)	−32.6 (−1.7%)	−88.2 (−4.5%)	−78.4 (−3.5%)	+45.3 (+2.1%)	−33.1 (−1.5%)
0.50	−59.6 (−4.4%)	−23.1 (−1.8%)	−82.7 (−6.1%)	−83.1 (−5.3%)	+71.4 (+4.8%)	−11.7 (−0.7%)
0.40	−55.6 (−6.0%)	−17.6 (−2.0%)	−73.2 (−7.9%)	−74.3 (−6.8%)	+81.6 (+8.1%)	+7.3 (+0.7%)
0.30	−41.2 (−6.8%)	−14.7 (−2.6%)	−55.9 (−9.3%)	−65.6 (−9.1%)	+72.5 (+11.0%)	+6.9 (+1.0%)
0.20	−28.2 (−7.7%)	−7.7 (−2.3%)	−35.9 (−9.8%)	−43.8 (−9.6%)	+56.3 (+13.6%)	+12.6 (+2.8%)
0.10	−16.7 (−8.2%)	−3.3 (−1.7%)	−20.0 (−9.7%)	−21.3 (−8.2%)	+43.2 (+18.1%)	+21.9 (+8.4%)

Table 5
Means (in mm^3) and coefficients of variation (CoV) of the change in T_1w lesion volume from baseline to 2 years, and rank correlations between change in black hole volume and change in MSFC, computed using the unpaired, registration-only and registration + T_2w mask averaging methods. The variable *l* indicates the maximum intensity, relative to NAWM and CSF, used to define the black holes, with $l=0.90$ being closest to NAWM and therefore the most inclusive. A value of $l=0.80$ corresponds to the traditional black hole definition of including all visually hypointense voxels. There is a clear increase in correlation strength when using the paired methods, especially T_2w mask averaging. For the correlation, *indicates $p<0.05$, **indicates $p<0.01$, ***indicates $p<0.001$ and ****indicates $p<0.0001$.

	Unpaired		Registration Only		Registration + T_2w mask averaging	
l	Vol. change (CoV)	Spearman correlation	Vol. change (CoV)	Spearman correlation	Vol. change (CoV)	Spearman correlation
0.90	920.82 (2.53)	−0.100	850.52 (2.53)	−0.064	542.74 (2.32)	−0.132*
0.80	623.51 (2.89)	−0.104	584.66 (2.83)	−0.066	518.73 (2.26)	−0.156*
0.70	415.21 (3.46)	−0.094	395.97 (3.31)	−0.091	430.49 (2.49)	−0.181**
0.60	292.6 (4.11)	−0.125	269.86 (4.04)	−0.119	347.69 (2.81)	−0.210***
0.50	215.99 (4.85)	−0.112	192.49 (4.77)	−0.133*	286.93 (3.16)	−0.226***
0.40	159.39 (5.56)	−0.121	140.67 (5.37)	−0.175**	239.95 (3.52)	−0.291****
0.30	119.98 (5.51)	−0.148*	95.51 (5.89)	−0.165**	182.74 (3.77)	−0.303****
0.20	90.38 (4.97)	−0.119	74.83 (5.46)	−0.145*	138.85 (3.97)	−0.300****
0.10	55.45 (4.80)	−0.081	50.85 (4.88)	−0.092	97.27 (4.00)	−0.203**

correlation, computed at $l=0.90$ (−0.132, $p=0.039$) is only slightly lower than the highest values observed with the unpaired (−0.148, $p=0.020$) and registration-only (−0.175, $p=0.006$) methods.

4. Discussion

The goal of our experiment was to study the impact of paired analysis for reducing the measurement noise in the longitudinal volumetric analysis of BHs. The hypothesis was that a reduction in the variability in partial volume and T_2w lesion segmentation across time would reveal the more pathologically relevant BH differences between time points and consequently improve the longitudinal correlations to disability.

4.1. Effect of registration on black hole volume

The effect of registration on the measurements in BH volume can be readily explained. The image resampling performed during registration blurs the images, and reduces the contrast between the BHs and surrounding NAWM. The blurring causes some NAWM pixels to become slightly hypointense, resulting in a slight expansion of the lesion boundaries. The blurring also shifts the intensity distribution of the lesions toward the higher intensities. The overall effect is that the BHs computed at the most inclusive intensity level ($l=0.90$) have slightly larger volume, while the volumes computed at the other levels are comparable or lower (up to −9.6% for the darker voxels). The symmetric registration affected the baseline and two-year volumes similarly, so the longitudinal changes computed in the registered images are similar or slightly lower than those in the unpaired analysis.

4.2. Effect of T_2w mask averaging on black hole volume

The effect of T_2w mask averaging on BH volume is more complex than for registration, but can also be explained with a number of observations. Mask averaging decreased the baseline volumes slightly (−1.7% to −2.6%) across all intensity levels compared to the registration-only method. The effect on the baseline volumes is minor because mask averaging mostly adds to the baseline masks in such a way that the additional regions largely include NAWM on the baseline scans and therefore add little to BH volume. The small decreases are likely attributable to areas of inflammation on the baseline scans that resolved before the follow-up or minor misregistration between the baseline and follow-up masks. In such areas, if the baseline T_2w mask value is less than 50% to begin with, slight underestimation of the contained BHs can occur.

The T_2w mask averaging method had a larger and different effect on the two-year volumes; while registration + mask averaging also produced lower values than the registration-only method for $l=$ 0.90 and 0.80, it computed higher volumes for $l\leq0.60$. The decreases for the higher values of *l* are likely due to underestimation in regions where there is a large number of newly developed lesion voxels in the follow-up scans. In such cases, the increase in the T_2w mask may be underestimated due to averaging, and the BHs within may also be underestimated, especially those voxels that are only slightly hypointense because they generally represent more recent damage and therefore are less likely to be in the baseline mask. In contrast, the darker BH voxels extracted by the lower values of *l* are robust to this effect because they generally represent more permanent damage and therefore are more likely to be found where there is good correspondence between the baseline and follow-up masks. For $l\leq0.60$, the increase in the two-year BH volumes caused by mask averaging can be explained by the fact that the cores of many T_2w lesions can shift in signal toward that of CSF as they evolve, becoming progressively darker on PDw and fluid-attenuated inversion recovery images (Guttmann et al., 1995; Rovaris et al., 1999). At some point, a T_2w lesion core can attain an intensity similar to that of NAWM, and if any part of the core lies near the lesion boundary, the T_2w lesion mask can be underestimated. Fig. 2 (top row, second image) shows an unusually dark core that results in a poor segmentation for that T_2w lesion (top row, last image). Fig. 3 illustrates a more typical example. While this type of error is usually small relative to the size of the T_2w lesion, it can also reduce the BH volume measurements.

Because the most hypointense BH voxels are typically inside the dark T_2w lesion core, and because these voxels are generally small in number, even a small segmentation error in the T_2w lesion core can induce a significant relative error in BH volume at the lower levels of l. Mask averaging reduces the underestimation because the baseline T_2w lesions typically have a less developed dark core, and can help re-establish the T_2w mask in the core region (Fig. 2, second row, last image), if the lesion is visible at baseline.

4.3. Improvements in longitudinal correlation

The improvements in longitudinal correlation to the MSFC resulting from paired analysis are characterized by an increase in the number of intensity levels for which a statistically significant correlation is found (from one for the unpaired method to four and nine for the registration-only and registration + mask averaging methods, respectively), as well as increases in the magnitudes of the correlations at each given level (from −0.148 at $l = 0.30$ for the unpaired method to −0.165 and −0.303 for the registration-only and registration + mask averaging methods, respectively). The improvements in longitudinal correlation from registration are likely due to the equalization of partial volume effects across the two time points. Even though the gains are modest, registration is a simple procedure and appears to be worth the effort. Comparing the two paired analysis methods, registration + mask averaging has a much greater impact than registration alone. The correlations between BH volume change computed by registration + mask averaging are higher for all intensity levels than the other methods, which is likely attributable to the removal of longitudinal variability in the search space formed by the T_2w masks, thereby increasing the ratio of true pathological change to measurement noise. In addition, for the lower values of l, mask averaging increased the accuracy of the darker BHs, which may represent greater injury, in the follow-up scans. Although the T_2w lesion segmentation has good accuracy, having attained a Dice coefficient (Dice, 1945) of 80% and cross-sectional rank correlation close to 1.0 as compared to a gold standard (McAusland et al., 2010), there is apparently still enough longitudinal variability to confound BH measurements when the T_2w lesion masks are applied to the T_1w scans for each time point individually. The results indicate that removing the variability of T_2w lesion segmentation by averaging the baseline and follow-up masks potentially improves the sensitivity of BH volume measurement to those changes that have the most clinical impact.

4.4. Cross-sectional vs. longitudinal trade-offs

In proposing a paired analysis, the anticipated trade-off was that improved longitudinal consistency in the segmentations would come at the expense of increased error in the cross-sectional segmentations, and that any increases in the longitudinal clinical correlations would be accompanied by reduced cross-sectional correlations. As discussed above, mask averaging results in a trade-off in the follow-up scans by underestimating the slightly hypointense areas while improving accuracy in the darker regions. Consequently, the volume changes computed by registration + mask averaging for $l = 0.90$ and 0.80 are lower than that computed by the other two methods, while the volume changes are higher for $l \leq 0.70$. However, the improvements in longitudinal clinical correlations for $l = 0.90$ and 0.80 strongly suggest that even though the volume changes may be generally underestimated, the ranking of the changes is more clinically relevant. In addition, while the longitudinal correlations have increased across all intensity levels, the cross-sectional correlations for the paired methods are very similar in magnitude and statistical significance to those produced by the unpaired method, which indicates that the differences in segmentation between the three methods are not large enough to cause many cross-sectional rank changes in volume in this patient sample. However, given that underestimation of some new lesions has been observed, it seems prudent to reserve the paired analysis for computing longitudinal volume changes, and use the unpaired method when cross-sectional segmentation accuracy is important.

4.5. Effect of intensity variations

The analysis of clinical correlations to BH volume at multiple levels of intensity, as done cross-sectionally in a previous study (Tam et al., 2011), was performed longitudinally in this study. This multi-level intensity analysis is a departure from the traditional approach of including all visually hypointense voxels, and has again proven useful as the magnitudes and statistical significance of both the cross-sectional and longitudinal correlations are seen to vary widely with the intensity threshold l. In the previous study, the strongest cross-sectional correlation was found at a low value of $l = 0.30$, suggesting that the darkest BHs had the greatest clinical impact, but in the current study the strongest cross-sectional correlations are found at $l = 0.90$. While the disagreement is somewhat disappointing, it is not completely unexpected as there are a number of key differences between the two studies, such as the size of the patient sample (24 vs. 247), clinical measure used (MSFC vs. Extended Disability Status Scale (Kurtzke, 1983)), and clinical status of the patients (mixed population of relapsing–remitting and SPMS patients vs. all SPMS). While the previous study did not have longitudinal data, the current study produced longitudinal correlations that are the strongest at $l = 0.30$ or 0.40 for all three computation methods, again suggesting greater clinical relevance for the darker BH voxels. While the results indicate that multi-level intensity analysis can enhance the understanding of the relationship between intensity variations and clinical correlations, to put the work in a broader context it may help to focus on the single intensity level that corresponds to the traditional approach of including all visually hypointense areas. In our algorithm an intensity level of $l = 0.80$ corresponds to the traditional BH definition. In this case, the registration + mask averaging method still shows a much improved correlation (−0.156, $p = 0.014$) over the registration-only (−0.066, $p = 0.305$) and unpaired (−0.104, $p = 0.103$) methods.

4.6. Limitations of study

While the improvements in longitudinal correlations are encouraging, further work is required to fully understand the benefits and limitations of paired analysis. We have so far only tested the idea on our own method for measuring BH volume. It would be important to validate the approach on other segmentation algorithms.

We have so far ignored scanner upgrades, which can cause longitudinal intensity changes that can affect BH classification. Our records indicate that 14% of the follow-up scans were acquired after a hardware upgrade and 37% were acquired after a software upgrade. Our experience is that software upgrades usually have a minor effect on the basic structural MR sequences such as the ones used in this study. Also, for all upgrades, a radiologist compared the scans acquired before and after for each site, and suggested parameter adjustments to minimize differences if required. However, the impact of upgrades on the computerized measurements should be more rigorously studied. In addition, changes in acquisition methods may enhance or offset the benefits of paired analysis and should be investigated. For example, high-resolution isotropic 3D MR sequences can reduce partial volume and improve lesion contrast over their conventional 2D spin-echo counterparts (Moraal et al., 2008), and can potentially further improve clinical correlations compared to our current results.

Our key assumption in using clinical correlation as the main outcome measure was that a stronger correlation meant that the extracted

BH voxels had greater pathological relevance, but T_1w lesion volume is a greatly simplified marker for MS pathology. The relationship between MS pathology and clinical status is unclear, especially when only structural MRI is used, and quantitative MR techniques that have greater pathological specificity such as diffusion tensor imaging, magnetization transfer ratio (MTR), and T_1 and T_2 relaxation maps (MacKay et al., 2009) are important for validating our assumption, and may produce stronger clinical correlations in the long term. In particular, the MTR of lesions has been identified as one of the most promising imaging biomarkers for MS clinical trials (Barkhof et al., 2009), and T_2-derived measures of myelin content have been shown to have high pathological relevance (Laule et al., 2006). However, the practicality of these techniques is still currently limited by technical difficulties in acquisition and analysis and they have yet to produce clinical correlations that match those of BHs (Poloni et al., 2011).

5. Conclusions

In conclusion, we have performed a first study on comparing unpaired and paired analysis of BH volume change in MS, and have shown that paired analysis, in particular T_2w mask averaging, can make strong improvements in clinical correlations. The proposed paired analysis methods can be implemented as relatively simple enhancements to most existing BH segmentation algorithms, but the results suggest that their impact can be powerful. Even though the study also revealed some limitations of the mask averaging method, overall the paired analysis approach appears very promising and warrants further investigation, especially when expanded to other patient populations and clinical measures.

References

Sahraian, M.A., Radue, E.-W., Haller, S., Kappos, L., 2010. Black holes in multiple sclerosis: definition, evolution, and clinical correlations. Acta Neurologica Scandinavica 122, 1–8.

Neema, M., Stankiewicz, J., Arora, A., Guss, Z.D., Bakshi, R., 2007. MRI in multiple sclerosis: what's inside the toolbox? Neurotherapeutics 4, 602–617.

van den Elskamp, I.J., Lembcke, J., Dattola, V., Beckmann, K., Pohl, C., Hong, W., Sandbrink, R., Wagner, K., Knol, D.L., Uitdehaag, B., Barkhof, F., 2008. Persistent T1 hypointensity as an MRI marker for treatment efficacy in multiple sclerosis. Multiple Sclerosis 14, 764–769.

Barkhof, F., Calabresi, P.A., Miller, D.H., Reingold, S.C., 2009. Imaging outcomes for neuroprotection and repair in multiple sclerosis trials. Nature Reviews. Neurology 5, 256–266.

Naismith, R.T., Cross, A.H., 2005. Multiple sclerosis and black holes: connecting the pixels. Archives of Neurology 62, 1666–1668.

Horsfield, M.A., Bakshi, R., Rovaris, M., Rocca, M.A., Dandamudi, V.S.R., Valsasina, P., Judica, E., Lucchini, F., Guttmann, C.R.G., Sormani, M.P., Filippi, M., 2007. Incorporating domain knowledge into the fuzzy connectedness framework: application to brain lesion volume estimation in multiple sclerosis. IEEE Transactions on Medical Imaging 26, 1670–1680.

Zhao, G., Li, D.K.B., Paty, D., 2000. MRI in multiple sclerosis. In: Mazziotta, J.C., Toga, A.W., Frackowiak, R.S.J. (Eds.), Brain Mapping: The Disorders. Academic Press, San Diego, pp. 357–381.

van Walderveen, M.A., Kamphorst, W., Scheltens, P., van Waesberghe, J.H., Ravid, R., Valk, J., Polman, C.H., Barkhof, F., 1998. Histopathologic correlate of hypointense lesions on T1-weighted spin-echo MRI in multiple sclerosis. Neurology 50, 1282–1288.

Barkhof, F., Bruck, W., De Groot, C.J.A., Bergers, E., Hulshof, S., Geurts, J., Polman, C.H., van der Valk, P., 2003. Remyelinated lesions in multiple sclerosis: magnetic resonance image appearance. Archives of Neurology 60, 1073–1081.

Riva, M., Ikonomidou, V.N., Ostuni, J.J., van Gelderen, P., Auh, S., Ohayon, J.M., Tovar-Moll, F., Richert, N.D., Duyn, J.H., Bagnato, F., 2009. Tissue-specific imaging is a robust methodology to differentiate in vivo T1 black holes with advanced multiple sclerosis-induced damage. AJNR. American Journal of Neuroradiology 30, 1394–1401.

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, 467–474.

Wu, Y., Warfield, S.K., Tan, I.L., Wells III, W.M., Meier, D.S., van Schijndel, R.A., Barkhof, F., Guttmann, C.R.G., 2006. Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. NeuroImage 32, 1205–1215.

Fischer, J.S., Rudick, R.A., Cutter, G.R., Reingold, S.C., 1999. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. Multiple Sclerosis 5, 244–250.

Tam, R.C., Traboulsee, A., Riddehough, A., Sheikhzadeh, F., Li, D.K.B., 2011. The impact of intensity variations in T1-hypointense lesions on clinical correlations in multiple sclerosis. Multiple Sclerosis Journal 17, 949–957.

McAusland, J., Tam, R.C., Wong, E., Riddehough, A., Li, D.K.B., 2010. Optimizing the use of radiologist seed points for improved multiple sclerosis lesion segmentation. IEEE Transactions on Biomedical Engineering 57, 2689–2698.

Filippi, M., Falini, A., Arnold, D.L., Fazekas, F., Gonen, O., Simon, J.H., Dousset, V., Savoiardo, M., Wolinsky, J.S., 2005. Magnetic resonance techniques for the in vivo assessment of multiple sclerosis pathology: consensus report of the white matter study group. Journal of Magnetic Resonance Imaging 21, 669–675.

Jones, C., Wong, E., 2002. A multi-scale application of the N3 method for intensity correction of MR images. Proceedings of SPIE, vol. 4684, pp. 1123–1129.

Smith, S.M., 2002. Fast robust automated brain extraction. Human Brain Mapping 17, 143–155.

Parzen, E., 1962. On estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076.

Caselles, V., Kimmel, R., Sapiro, G., 1997. Geodesic active contours. International Journal of Computer Vision 22, 61–79.

McAusland, J., Wong, E., Riddehough, A., Li, D.K.B., 2004. A modified fuzzy clustering method for modelling partial volume effects in MRI data. Proceedings of the 12th Annual Meeting of ISMRM.

Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A., 2003. Mutual-information-based registration of medical images: a survey. IEEE Transactions on Medical Imaging 22, 986–1004.

Meijering, E.H., Niessen, W.J., Viergever, M.A., 2001. Quantitative evaluation of convolution-based methods for medical image interpolation. Medical Image Analysis 5, 111–126.

Guttmann, C.R., Ahn, S.S., Hsu, L., Kikinis, R., Jolesz, F.A., 1995. The evolution of multiple sclerosis lesions on serial MR. AJNR. American Journal of Neuroradiology 16, 1481–1491.

Rovaris, M., Comi, G., Rocca, M.A., Cercignani, M., Colombo, B., Santuccio, G., Filippi, M., 1999. Relevance of hypointense lesions on fast fluid-attenuated inversion recovery MR images as a marker of disease severity in cases of multiple sclerosis. AJNR. American Journal of Neuroradiology 20, 813–820.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Kurtzke, J.F., 1983. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33, 1444–1452.

Moraal, B., Roosendaal, S.D., Pouwels, P.J.W., Vrenken, H., van Schijndel, R.A., Meier, D.S., Guttmann, C.R.G., Geurts, J.J.G., Barkhof, F., 2008. Multi-contrast, isotropic, single-slab 3D MR imaging in multiple sclerosis. European Radiology 18, 2311–2320.

MacKay, A.L., Vavasour, I.M., Rauscher, A., Kolind, S.H., Mädler, B., Moore, G.R.W., Traboulsee, A.L., Li, D.K.B., Laule, C., 2009. MR relaxation in multiple sclerosis. Neuroimaging Clinics of North America 19, 1–26.

Laule, C., Leung, E., Li, D.K., Traboulsee, A.L., Paty, D.W., MacKay, A.L., Moore, G.R., 2006. Myelin water imaging in multiple sclerosis: quantitative correlations with histopathology. Multiple Sclerosis 12, 747–753.

Poloni, G., Minagar, A., Haacke, E.M., Zivadinov, R., 2011. Recent developments in imaging of multiple sclerosis. The Neurologist 17, 185–204.