NeuroImage: Clinical 3 (2013) 171–179

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Quantification of multiple-sclerosis-related brain atrophy in two heterogeneous MRI datasets using mixed-effects modeling ☆

Blake C. Jones [a,b,1], Govind Nair [a], Colin D. Shea [a], Ciprian M. Crainiceanu [c], Irene C.M. Cortese [a], Daniel S. Reich [a,c,*]

[a] Translational Neuroradiology Unit, National Institute of Neurological Disorders and Stroke, NIH, 10 Center Drive MSC 1400, Building 10 Room 5C103, Bethesda, MD 20892, USA
[b] Case Western Reserve University School of Medicine, 10900 Euclid Ave., Cleveland, OH 44106, USA
[c] Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, E3636, Baltimore, MD 21205, USA

ARTICLE INFO

Article history:
Received 23 March 2013
Received in revised form 31 July 2013
Accepted 1 August 2013
Available online xxxx

Keywords:
Multiple sclerosis
MRI
Brain atrophy
Mixed-effects model
Heterogeneous data

ABSTRACT

Brain atrophy, measured by MRI, has been proposed as a useful surrogate marker for disease progression in multiple sclerosis (MS). However, it is conventionally assumed that the accurate quantification of brain atrophy is made difficult, if not impossible, by changes in the parameters of the MRI acquisition, which are almost inevitable over the course of a longitudinal study since MRI technology changes rapidly. This state of affairs can negatively affect clinical trial design and limit the use of historical data. Here, we investigate whether we can coherently estimate brain atrophy rates in a heterogeneous MS sample via linear mixed-effects multivariable regression, incorporating three critical assumptions: (1) using age at time of scanning, rather than time since baseline, as the regressor of interest; (2) scanning individuals with a variety of techniques; and (3) introducing a simple additive correction for major differences in MRI protocol. We fit the model to several measures of brain volume as the outcome in two MS populations: 1123 scans from 195 cases acquired for over approximately 7 years in two natural history protocols (Cohort 1), and 1331 scans from 69 cases seen for over 11 years who were primarily treated with two specific MS disease-modifying therapies (Cohort 2). We compared the mixed-effects model with additive correction for MRI acquisition parameters to a model fit without this correction and performed sample-size calculations to provide an estimate of the number of participants in an MS clinical trial that might be required to see a therapeutic effect of treatment using the approach described here. The results show that without the additive correction for T1-weighted protocol parameters, atrophy was underestimated and subject-specific estimates were more narrowly distributed about the population mean. Ventricular CSF is the most consistently estimated brain volume, with a mean of 2.8%/year increase in Cohort 1 and 4.4%/year increase in Cohort 2. An interesting observation was that gray matter volume decreased and white matter volume remained essentially unchanged in both cohorts, suggesting that changes in ventricular CSF volume are a surrogate for changes in gray matter volume. In conclusion, the mixed-effects modeling framework presented here allows effective use of heterogeneously acquired and historical data in the study of brain atrophy in MS, potentially simplifying the design of future single- and multi-site clinical trials and natural history studies.

1. Introduction

Multiple sclerosis (MS) is an immune-mediated disease of the CNS that is characterized by focal demyelinating lesions and causing physical and cognitive impairment. Although these focal lesions are the most striking feature of the disease when MS patients are studied with magnetic resonance imaging (MRI), brain lesion volume per se is not strongly correlated with disability (Filippi et al., 1995; Furby et al., 2010; Kappos et al., 1999). Additionally, patients often continue to have progressive symptoms while developing few new lesions.

One possible explanation for progressive disease in MS despite a paucity of new lesions is abnormally rapid brain atrophy, which has been noted in many studies that have examined MS with MRI (for review, see Giorgio et al., 2008). Atrophy has its origin in neuronal and

Abbreviations: MRI, magnetic resonance imaging; MS, multiple sclerosis; vCSF, ventricular CSF; sGM, supratentorial gray matter; sWM, supratentorial white matter; sTot, supratentorial total volume; FSPGR, fast spoiled gradient echo; MPRAGE, magnetization-prepared rapid acquisition of gradient echoes; EDSS, Kurtzke Expanded Disability Status Scale.

* Corresponding author at: Translational Neuroradiology Unit, National Institute of Neurological Disorders and Stroke, NIH, 10 Center Drive MSC 1400, Building 10 Room 5C103, Bethesda, MD 20892, USA. Tel.: +1 301 496 1801.
E-mail addresses: bcj20@case.edu (B.C. Jones), govind.bhagavatheeshwaran@nih.gov (G. Nair), sheacd@ninds.nih.gov (C.D. Shea), ccrainic@jhsph.edu (C.M. Crainiceanu), corteseir@ninds.nih.gov (I.C.M. Cortese), daniel.reich@nih.gov (D.S. Reich).
[1] Present address: Loyola University Medical Center. 2160 South First Avenue, Maywood, IL 60153, USA.

2213-1582/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.nicl.2013.08.001

glial degeneration, primarily resulting in gray matter loss (Fisher et al., 2008; Fisniku et al., 2008; Shiee et al., 2012). Gray matter volume has been shown to be an independent predictor of disability in all MS subtypes (Roosendaal et al., 2011). Therapies such as interferon beta-1b (Molyneux et al., 2000), glatiramer acetate (Rovaris et al., 2001), and natalizumab (Miller et al., 2007), which are effective at preventing new lesions, are ineffective at fully correcting atrophy rate (for review of the effect of therapeutics on atrophy, see Zivadinov et al. (2008)). Thus, atrophy rate has been postulated as a surrogate marker for disease progression and has been used as an outcome measure in many trials (several previously mentioned, as well as Filippi et al. (2004)).

An issue with using brain atrophy as a surrogate for disease progression is the inherent difficulty in measuring it. Most studies use normalized measures of brain volume that fall primarily into two categories. Longitudinal methods, such as SIENA, register two images from the same individual and look for changes in the brain surface (Smith et al., 2002). Cross-sectional methods relate the volume of interest to a normalization volume, such as the skull or intracranial volume, and track changes in the relative size of brain structures over time (Chard et al., 2002). Both types of method have been valuable in the study of atrophy in MS but require, or perform best with, MRI data acquired in a homogeneous fashion with respect to hardware, protocol, and acquisition parameters. The reason is that these methods are analogous to studying the atrophy rate of each subject separately and averaging the results to generate an atrophy rate for the population of subjects. A spurious change in apparent brain volume induced by a change in acquisition parameters in one or more scans on a particular subject will produce an inaccurate measurement of atrophy rate in that subject. A "second stage," population-level analysis will then produce a flawed population estimate of atrophy rate. This technical constraint limits study duration and sample size, since imaging technology is in constant flux and different centers often use different equipment and protocols, and means that brain atrophy is often estimated from relatively short-term data. However, short-term biological variability in brain volume is well-known in MS, resulting from variations in factors such as hydration status and the effect of anti-inflammatory therapy known as pseudoatrophy, which can result in loss of brain volume without actual loss of brain tissue upon the initiation of treatment (Fox et al., 2005; Mellanby and Reveley, 1982; Rao et al., 2002; Sampat et al., 2010; Zivadinov et al., 2008). Compounding the problem is that the absolute change in brain volume that represents real atrophy is small relative to this biological variability (De Stefano et al., 2010; Rao et al., 2002). The result — that follow-up must essentially start over with every new trial or improvement in MRI equipment or technique — is especially difficult when good quality historical control data from MS trials exist (Shirani et al., 2012).

In the analysis of longitudinal data, linear mixed-effects multivariable regression analysis enjoys several well-described statistical advantages over the described "two-stage" or so-called "NIH method" (Cnaan et al., 1997) (for textbook treatment, see Fitzmaurice et al. (2011)). Mixed-effects regression enables the generation of both population-level and subject-specific atrophy rates, permits the use of data from subjects with few data points, and also allows the quantification of the effects on atrophy of variables such as therapy and lesion load at the population level. In fact, mixed-effects models have been applied to atrophy rates generated using the aforementioned methods to calculate sample sizes for trials (Anderson et al., 2007).

The goal of this study was to investigate whether longitudinal data from multiple scanning protocols can be reasonably integrated into a coherent estimate of individual- and population-level brain volume changes in MS. We hypothesized that applying linear mixed-effects regression to absolute brain volumes, rather than to atrophy rates generated separately for each patient, would allow us to correct for the effect of MRI acquisition differences at the population level using an additive factor. To accomplish this, we postulate that regression should be performed against age at the time of scanning (rather than time elapsed since the first scan, with age as covariate), and also that scans with different protocols be performed on individual subjects. We show that this approach can generate reasonable results for atrophy rate in two populations of MS patients with intra-subject variation in MRI acquisition parameters — Cohort 1, a collection of heterogeneous longitudinal data from two MS natural history protocols, including scans performed off and on disease-modifying therapy; and Cohort 2, derived from a retrospective analysis of data obtained from MS cases seen at our center over an 11-year period who were predominantly treated with daclizumab, an anti-CD25 monoclonal antibody, and interferon beta. An important implication of this work is that it opens the doors for comparison of historical data to data more recently acquired, particularly if individual subjects were studied with a spectrum of protocols, thereby decreasing the impact of small sample sizes and short-term follow-up durations on studies of brain atrophy.

2. Materials and methods

2.1. Cohort 1

This cohort includes a convenience sample of MS cases enrolled in our center between October 2005 and September 2011, who were being evaluated in the context of two ongoing natural history protocols (the last scan in this cohort was performed in December 2011). We initially analyzed 1491 MRIs from 373 cases using our automated image-processing software (see below). Of these, we excluded 111 cases with only one MRI, as we were only interested in longitudinal changes in brain volume. Of the remaining 262 cases, we excluded 3 because our software failed to provide reasonable estimates of lesion volume from their images. We also removed 64 cases because their final diagnosis was not MS according to 2005 and 2010 McDonald criteria (Polman et al., 2005, 2011). A final pool of 1123 MRIs from 195 cases remained for the analysis. Of these scans, 23 were later removed because they had extreme ventricular sizes that precluded reliable quantification of brain volume (see Section 2.4). Table 1 contains a summary of demographic and clinical characteristics; a summary of MRI acquisition parameters may be found in Table 2. Subjects in Cohort 1 were treated with a variety of disease modifying therapies, including interferon beta 1a and 1b, glatiramer acetate, daclizumab, natalizumab, idebenone, mitoxantrone, and fingolimod; 773 scans were from subjects treated with any disease-modifying agent during the observation period, while 327 were from those never treated during this period. Modeling of individual treatment effects was beyond the scope of this work. Differences in follow-up length were not considered in our analysis, as a multivariable regression of age, sex, EDSS, and disease type on follow-up length (results not reported) yielded no association between any of these factors and follow-up length.

Table 1
Relevant patient sample characteristics for both Cohort 1, a convenience sample of MS patients from two natural history protocols, and Cohort 2, patients whose data were used in a post hoc analysis of the therapeutic effect of daclizumab in MS. MS: RR, relapsing–remitting; PP, primary progressive; SP, secondary progressive; CIS, clinically isolated syndrome.

Dataset	Cohort 1	Cohort 2
N (by sex)	116 F, 79 M	44 F, 25 M
Number of scans	1123	1331
Mean age (range)	42 years (17–68 years)	38 years (18–60 years)
Median scans per person (range)	4 (2–27)	15 (2–55)
Mean total follow-up time (range)	1.2 years (21 days–5.2 years)	5.9 years (1.1–10.5 years)
Disease type	RRMS: 142; PPMS: 34; SPMS: 12; CIS: 7	All RRMS
Median EDSS (range)	1.5 (0–8.5)	1.5 (0–6.5)

Table 2
MRI acquisition parameters. Nominal in-plane resolution was approximately 1 × 1 mm for all scans. T1WP, T1-weighted protocol; Vol, volume coil; 8-ch, 8-channel head coil; TR, repetition time; TE, echo time; TI, inversion time; F, FSPGR (fast spoiled gradient echo); M, MPRAGE (magnetization-prepared rapid acquisition of gradient echoes).

Cohort 1											Cohort 2			
T1WP:	1	2	3	4	5	6	7	8	9	10	1	2	3	4
Number of scans	5	335	2	1	68	305	5	289	30	83	188	996	38	110
Scanner manufacturer	GE	GE	GE	GE	GE	GE	GE	GE	GE	Philips	GE	GE	GE	GE
Field strength (T)	1.5	1.5	1.5	1.5	1.5	1.5	1.5	3	3	3	1.5	1.5	1.5	1.5
Receive coil	Vol	Vol	Vol	Vol	8-ch	8-ch	8-ch	8-ch	8-ch	8-ch	Vol	Vol	8-ch	8-ch
TR (ms)	12	9–10	7	8	10	9	8	9	8–9	7	~9	~9	~10	~9
TE (ms)	5	2–3	3	3	3	3.5	3	3.5	3	3	~2	~2	~3	~3.5
TI (ms)	None	None	400	750	None	450	750	450	725	900	None	None	None	450
Flip angle (deg)	20	20	12	16	20	13	16	13	6	9	20	20	20	13
Slice thickness (mm)	1.2	1.4	1.5	1	1.4	1.5	1	1	1	1	1.3	1.4	1.4	1.5
Acquisition protocol	F	F	F	F	F	F	F	F	M	M	F	F	F	F

2.2. Cohort 2

Cohort 2 consisted of MS cases whose data were used in a post hoc analysis of the therapeutic effect of daclizumab (Borges et al., 2013). Seven of the 69 cases in Cohort 2 were also in Cohort 1 (comprising only 20 scans). Table 1 contains clinical and demographic characteristics of these data, while Table 2 shows the MRI acquisition parameters. There were 26 cases treated with daclizumab and 43 non-daclizumab cases; the effect of therapy was not considered here.

2.3. Image processing

Lesion-TOADS is a fully automated, topology-preserving brain segmentation algorithm based on image intensity and a statistical prior probability atlas that enables simultaneous tissue classification and MS lesion detection and compares favorably with other segmentation techniques (Shiee et al., 2010). The algorithm uses both a T1-weighted image and a T2-FLAIR image to generate a classifier image and accounts for the effect of lesions on signal intensity in the T1-weighted image. The output from lesion-TOADS includes brain volumes for each of the 17 tissue types it considers, as well as the lesion volume. The T1-weighted images are used for brain structure segmentation, whereas the T2-FLAIR images are used for lesion segmentation. In this study, our tissue types of interest were ventricular CSF (vCSF) as well as supratentorial gray matter (sGM), white matter (sWM), and total volume (sTot). Average lesion volume, calculated as described below, was also considered. sWM volume included both lesions and normal-appearing white matter.

Before applying lesion-TOADS, T1-weighted and T2-FLAIR images were inhomogeneity corrected using the N3 algorithm (Sled et al., 1998), and the chronologically middle image was rigidly registered to the MNI-152 space; this middle image was also used as the target for registration of all other images for that subject. A mean image was then computed by voxel-wise averaging across all of the subject's scans. The skull was removed automatically using SPECTRE (Carass et al., 2007), generating a brain-and-CSF mask for each individual scan. Each image was normalized by subtracting the mean whole brain intensity and dividing by its standard deviation (Shinohara et al., 2011), and the images were then re-registered to the mean image. Lastly, the histograms for each image were matched to the histogram of the mean image using the "Histogram Image Matching" algorithm in MIPAV (medical image processing, analysis, and visualization, http://mipav.cit.nih.gov). Lesion-TOADS was then applied to each image.

A single lesion volume was generated for each case by averaging all volumetric T1-weighted images and all T2-FLAIR images from that case and applying lesion-TOADS to the average images. In our experience, this produced a more reliable single estimate of overall lesion volume than averaging the lesion load estimates from individual scans, at the cost of losing the ability to track lesion volume over time and possibly inducing an overestimation of lesion volume (as lesions tend to accumulate over time).

2.4. Statistical methods

A linear multivariable mixed-effects regression model was applied separately to each tissue type of interest to quantify the rate of volume change:

$$Vol_{ij} = \beta_0 + \beta_1 * AGE_{ij} + \beta_2 * SEX_i + \beta_3 * LESION_i + \beta_4 * T1WP_{ij} + b_{0i} + b_{1i} * AGE_{ij} + \varepsilon_{ij}.$$

In this model, the predicted volume of interest for subject i at time j (Vol_{ij}) is determined by population ("fixed") effects, represented by $\beta_0 \ldots \beta_4$, and subject-specific ("random") effects, b_{0i} and b_{1i}. The average rate of change in the population is β_1, while the subject-specific difference in that rate is captured by b_{1i}. Sex and T1-weighted protocol (T1WP) were modeled as indicator variables. T2-FLAIR protocol is not included because the T2-FLAIR image is predominantly used by lesion-TOADS for delineation of lesions rather than of brain tissue structure. Differences in T1WP were captured in a descriptive "code" generated from the scanner manufacturer, magnetic field strength, receive coil, acquisition type (fast spoiled gradient echo [FSPGR] vs. magnetization-prepared rapid acquisition of gradient echoes [MPRAGE]), inversion time (TI), and slice thickness. Data are presented with this correction applied, so that model fits can be shown as straight lines. Age in years and lesion load (LESION) are continuous variables. The model was fitted once to vCSF volume data, and an empirical threshold for the residuals was established (absolute value of the residual >3.75 ml). Scans with data above this cutoff were found to be those with extreme ventricular sizes. These scans were removed (23 scans), and the model was refit for each structural volume of interest. The rationale for this is that lesion-TOADS most often fails on scans with extremely large or small ventricles due to its use of a statistical atlas for tissue types. The results of the model were compared to those obtained when fitting the same model but removing T1WP; a likelihood ratio test was performed to compare the two models. For vCSF volume in Cohort 1, results for fitting a model including a fixed effect for the square of age as well as a model including a fixed effect for ever treated vs. never treated during the observation period are also reported. Additionally, the effect of using age as the time variable was compared to structuring the model with time elapsed since first enrollment as the time variable and age as a covariate. For Cohort 2, longitudinal lesion load was not included in the previously reported modeling results (Borges et al., 2013).

The distribution of vCSF volumes in the population as a whole was found to be right-skewed in both datasets, so a natural logarithmic transformation was applied before model fitting. Therefore, atrophy rates for vCSF are given as %/year. A normal distribution reasonably approximated all of the supratentorial tissue measures, so no transformation

was applied to these data, and atrophy rates are reported as absolute rates in ml/year.

Statistical analysis was performed in R version 2.14.1 (R Development Core Team, 2011). The lme4 package version 0.999375-42 (Bates et al., 2011) was used for the fitting of mixed effects models using restricted maximum likelihood estimation. Graphics were produced using the ggplot2 package, version 0.9.1 (Wickham, 2009). In lieu of p values, the 95% confidence interval is reported for calculated values as appropriate, since many comparisons were performed and showing "statistical significance" of the results is not the goal of these analyses.

2.5. Sample size calculations

The sample sizes necessary in each arm of a hypothetical therapeutic trial to detect effects of 25%, 50%, and 75% of the therapy on the atrophy rate were calculated using the following formula (Diggle et al., 2002):

$$N = \frac{2}{\gamma^2}\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2\left(\sigma_{b,2}^2 + 2\frac{\sigma_\varepsilon^2}{t^2}\right)$$

where N represents the number of subjects per trial arm necessary to detect a change in atrophy rate γ, assuming variance of the slopes $\sigma_{b,2}^2$ and variance of the residuals ${\sigma_\varepsilon}^2$, both of which were estimated from our sample. Z-scores for $\alpha = 0.05$ and power $= 1 - \beta = 80\%$ were used. Results were calculated for a simplified trial with an initial scan and one follow-up scan; follow-up scan times of one and two years were compared.

3. Results

Fig. 1 shows representative T1-weighted and T2-FLAIR images from one case at a single time point, as well as the lesion-TOADS classifier mask associated with them.

Mixed-effects models with and without T1WP correction were fitted to the data collected from Cohort 1, a heterogeneous population of MS cases (see Materials and methods). The effect of including an additive correction factor based on a code that captures the variability associated with 6 parameters of the MRI acquisition is demonstrated in Fig. 2. The additive corrections for each T1WP code were calculated from the mixed-effects model fits and are constrained by the population-level effects of age, sex, and total lesion volume, as well as the clustering of data by subject. These correction factors were then explicitly added to, or subtracted from, the original measurements in order to create the T1WP-corrected values shown in the plots. Other acquisition types were corrected to protocol 1 (see Table 2). The mixed effects fits are shown with T1WP-correction (black lines) and without it (red lines).

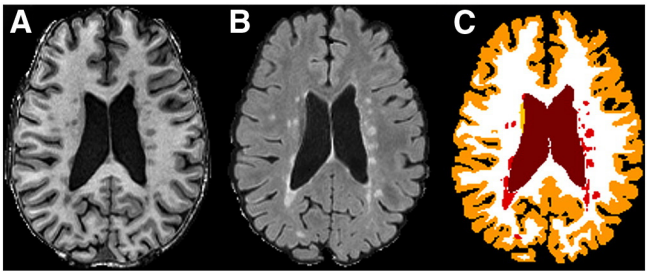

Fig. 1. Representative A) T1-weighted (FSPGR), B) T2-FLAIR, and C) lesion-TOADS classification images. The lesion-TOADS image shows classification of cortical gray matter (orange), white matter (white), ventricles (brown), and lesions (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The effect of T1WP correction is explored in two individual subjects in Fig. 3, which includes notation of the combinations of acquisition parameters for each scan (with reference to Table 2).

The population estimates of atrophy rate, as well as the variability in the subject-specific estimates as quantified by the standard deviation, are given in Table 3. Additionally, Table 3 provides results for each tissue type when the model includes time elapsed since enrollment/first scan as the time variable, with age as covariate. Note that likelihood ratio tests comparing the model results without and with the T1WP correction were all highly significant ($p < 2.2 \times 10^{-16}$), probably due to the high number of observations and relatively small number of terms in the model. Moreover, residual variance is substantially lower for the models that include the T1WP correction.

The results indicate that without the correction for T1WP, the model assigns a lower population rate of vCSF increase and a higher rate of sWM decrease relative to the results obtained with the T1WP correction. Stability over time of sWM volume, as determined by the T1WP-corrected model, is consistent with prior studies (Pirko et al., 2007; Shiee et al., 2012). Moreover, the uncorrected models yielded a narrower distribution of subject-specific atrophy rates around the population mean, whereas the T1WP-corrected model closely approximated individual subject trends, particularly for vCSF. Age and time elapsed both provide similar estimates of atrophy for vCSF when used as the time variable, although the variance in the estimate is larger when using time elapsed. For sGM and sWM, however, using age more strongly constrains the estimate of atrophy rate in the face of the high variability in these data. Residual variance is similar for these two classes of model.

A model including a fixed effect for the square of age reduced the utility of the age predictor while not adding great predictive value to the model for vCSF data. In particular, in the quadratic model the atrophy rate with age was 2.0%/year [95% CI: −1.4, 4.3] and with age^2 0.011%/year [−4.2, 4.4]; the SD of atrophy rates was 2.6%/year. Hence, we performed further analysis without the quadratic term. A fixed effect of ever vs. never treated with any disease-modifying agent during the observation period did not add predictive value to the model (regression coefficient for treatment: 15%/year [−1.3,35]) and was also excluded from further analyses.

Fig. 4 shows the mixed-effects fits with and without T1WP correction for the cases from Cohort 2. To facilitate comparison to Cohort 1, no treatment effect of daclizumab was considered, unlike in Borges et al. (2013), where daclizumab therapy was shown to reduce the rate of brain volume change relative to interferon beta. Population atrophy rate and within-subject variability estimates for Cohort 2 are found in Table 4. In this cohort, long follow-up and relatively low within-subject variability (due in part to the use of more similar acquisition protocols) led to reasonable individual- and population-level estimates for both vCSF and sTot. As in Cohort 1, atrophy was underestimated, and subject-specific estimates were more narrowly distributed about the population mean, in the absence of T1WP correction. Note that we did not separate the data into gray and white matter components for Cohort 2, because gray-white segmentation was not particularly reliable in this dataset, where most scans were acquired at 1.5 T with a volume receive coil (Borges et al., 2013).

To study the effect of the differing within-subject variability between vCSF and sGM in Cohort 1 on the ability to detect changes in atrophy rate, we performed calculations to estimate the sample sizes in each arm of therapeutic trials necessary to detect 25%, 50%, and 75% differences in atrophy rate. A simplified trial design of one initial scan and one follow-up scan at either one or two years was used (Table 5). The results suggest that about 18 times as many cases are required to detect changes in atrophy rate for sGM volume as compared to vCSF volume for the one-year trial, whereas about 10 times as many cases are necessary for the two-year trial. Note that these sample sizes are provided for illustration only, as it is highly unlikely that a clinical trial would be designed using a cohort and acquisition parameters similar to ours.

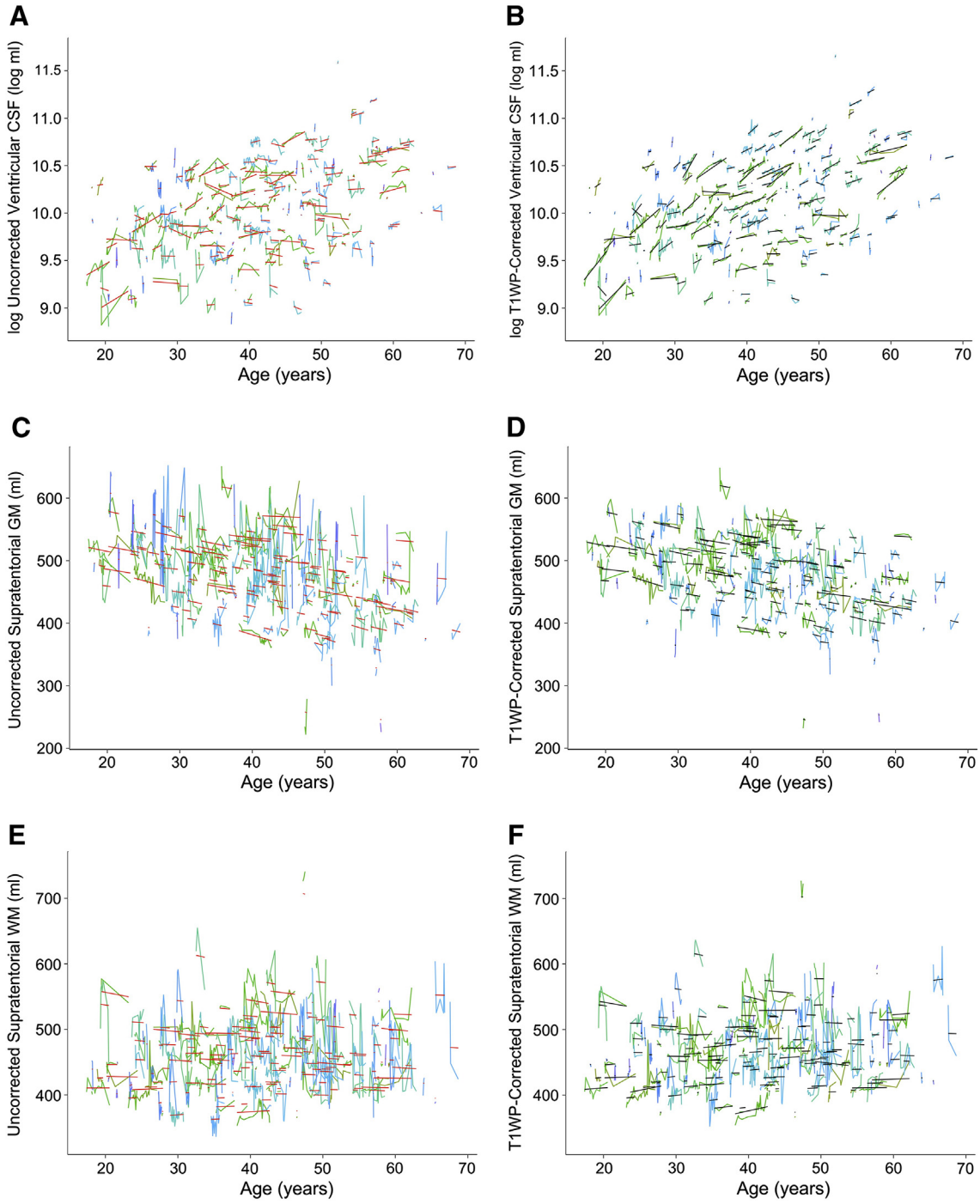

Fig. 2. The effect of T1-weighted protocol (T1WP) correction in 195 MS cases in Cohort 1 (natural history protocol; individual cases represented with different hues of blue and green). A) Log-transformed ventricular cerebrospinal fluid (CSF) volume and fits generated by the mixed effects model that did not incorporate T1WP correction (black lines) B) The fits (red lines) of the model with T1WP correction for ventricular CSF; the point-by-point data were also corrected for T1WP. C) and D) are analogous plots for supratentorial gray matter (GM) volume, while E) and F) show the same for supratentorial white matter (WM) volume. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

Brain atrophy in MS is a well-studied phenomenon that is thought to reflect neuroaxonal loss and that, as a result, has been postulated as a surrogate measure of disease progression. Our results demonstrate that mixed-effects modeling of absolute brain volumes can robustly allow the use of larger, more heterogeneous datasets to improve estimation of atrophy rates in MS populations. The methodology presented here may be further tested in large-scale, multi-site clinical trials.

Specifically, our framework incorporates: (1) explicit integration of age as the independent variable of interest, given clear population-wise (and essentially linear) associations of brain structure volumes with age; (2) simple statistical modeling of heterogeneous data acquisition protocols to reduce uninformative variability in the data, which is facilitated by acquiring data from individual subjects using a variety of protocols; and (3) generation of subject-specific rates of brain volume change.

The goal of this study was not to define the optimal model for integrating heterogeneous data into a single analysis. Nevertheless, it is

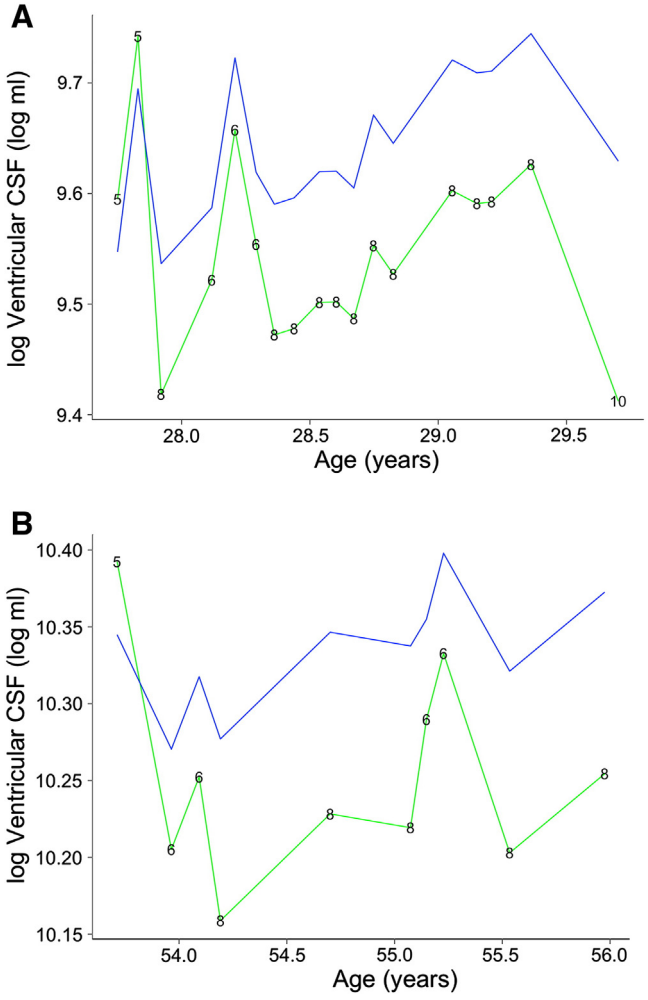

Fig. 3. The effect of T1-weighted protocol (T1WP) correction in A) a representative subject who underwent 18 scans for over 2 years with 4 different T1-weighted protocols and B) a second representative subject who underwent 10 scans in 2 years and 4 months. Plots show log-transformed ventricular cerebrospinal fluid (CSF) volume; the green line represents uncorrected data, while the blue line represents data with additive T1WP corrections applied. The number on each uncorrected point represents protocol number, as described in Table 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interesting to note that even the simple additive correction used here substantially reduces within-subject variability. In-depth modeling of the effects of individual scanning parameters is beyond the scope of this study but could certainly be incorporated to further reduce the variability and improve estimation of atrophy rates. Also beyond the scope of this study is determining the specific numbers of subjects and scans per protocol that would be necessary to apply the methods described here. Although the success of the method would certainly increase with the number of subjects and scans, in our experience simple inspection of the corrected volume trajectories should yield a good qualitative sense of the success or failure for a particular dataset. An additional refinement of the model that includes a fixed effect for the square of age in estimation of atrophy rates did not have increased predictive power, consistent with a linear atrophy rate over time as seen in several reports (Martola et al., 2010; Shirani et al., 2012). Further refinements of the model to test the effects of lesion load and disease subtype, as well as interactions of scanning protocol with time, could proceed under the same framework. Additionally, the effect of specific disease-modifying therapies could be investigated in much more detail. A simplistic ever vs. never treated model in Cohort 1, a convenience sample, did not increase predictive power, but applying a similar but more specific model in Cohort 2 yielded a clear effect of daclizumab therapy (Borges et al., 2013).

We note that the proposed framework can be applied to any scalar, continuous variable: any measured quantity derived from imaging or other data can be used as the outcome variable, as long as the variable (or a transform thereof) varies deterministically with age across the population. Our data suggest that using age as the time variable in data with relatively lower variability (e.g., vCSF) provides estimates similar to those obtained using time elapsed since enrollment (although using age does provide a narrower confidence interval even in this case). However, as variance in the data increases, as occurs for sGM and sWM volumes, using time elapsed provides too weak a constraint upon the high variability between scans to tease out the population-level atrophy rate.

These considerations make the mixed-effects modeling framework proposed here a very strong option in the population-level analysis of brain atrophy in MS, especially when long-term data are available. The model could even conceivably be applied in clinical practice to make inferences about individual patients whose historical images were collected with varying acquisition parameters — a very common situation, particularly at referral centers where patients bring prior MRI scans at the time of first evaluation. For example, a particular MS center might apply the model to determine a correction factor for brain volume for each scanner and protocol combination, which would allow the interpretation of brain volume changes on the subject level over long periods of time.

One interesting outcome of this study is the suggestion that vCSF volume can stand in for sGM volume, since sWM volume appears to change very little (if at all) over time. There is little doubt that primary vCSF expansion is unlikely in MS — rather, changes in vCSF are primarily due to the loss of sGM, the primary component of atrophy in MS (Fisher et al., 2008; Fisniku et al., 2008; Shiee et al., 2012). Of note, we did not attempt to separate different gray matter structures, such as cortical vs. deep gray matter, and we did not examine changes in infratentorial gray matter volume. These insights are supported by a recent study showing that lesion volume is correlated more strongly with global than regional atrophy, as well as a study describing the ability of lateral ventricle volume to differentiate stable from progressive patients (Antulov et al., 2011; Horakova et al., 2009).

Table 3
Estimates of atrophy rates, with associated 95% confidence intervals (CI) as well as standard deviations (SD) for subject-specific slopes, using the mixed-effects model with and without T1WP correction for ventricular cerebrospinal fluid (vCSF), supratentorial gray matter (sGM), and supratentorial white matter (sWM) volumes in Cohort 1. The residual variance of the model is also reported. The last row shows the calculated atrophy rates when time elapsed since enrollment is used as the time variable in the mixed-effects model with age as a covariate. T1WP, T1-weighted protocol.

Model	vCSF rate [95% CI]	SD of vCSF rates	Residual variance [(log-mm^3)^2]	sGM rate [95% CI]	SD of sGM rates	Residual variance	sWM rate [95% CI]	SD of sWM rates	Residual variance
T1WP correction	2.8%/year [2.1, 3.5]	±2.6%/year	1.6×10^{-3}	−2.1 ml/year [−2.7, −1.4]	±0.77 ml/year	190 ml^2	0.18 ml/year [−0.42, 0.78]	±0.70 ml/year	310 ml^2
No T1WP correction	0.72%/year [0.15, 1.3]	±1.7%/year	3.6×10^{-3}	−2.1 ml/year [−2.8, −1.5]	±0.67 ml/year	610 ml^2	−0.46 ml/year [−1.0, 0.11]	±0.55 ml/year	540 ml^2
T1WP correction, time since enrollment	2.9%/year [1.5, 4.2]	±4.8%/year	1.3×10^{-3}	0.47 ml/year [−2.0, 2.9]	±5.2 ml/year	160 ml^2	0.14 ml/year [−2.5, 2.8]	±5.6 ml/year	280 ml^2

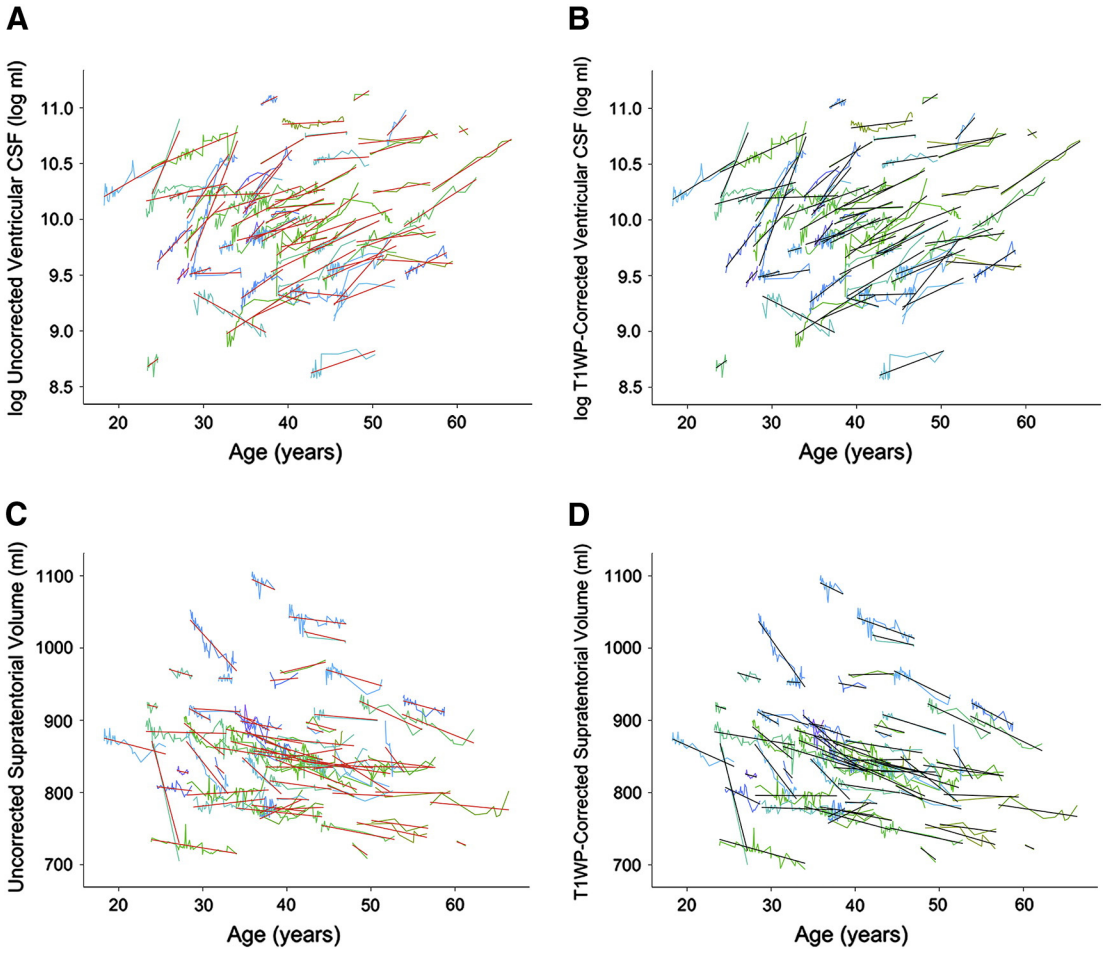

Fig. 4. The effect of T1-weighted protocol (T1WP) correction in 69 MS cases in Cohort 2 (cases used in a post-hoc analysis of daclizumab (Borges et al., 2013); individual cases represented with different hues of blue and green). A) Log-transformed ventricular cerebrospinal fluid (CSF) volume and fits generated by the mixed effects model that did not incorporate T1WP correction (black lines). B) The fits (red lines) of the model with T1WP correction for ventricular CSF; the point-by-point data were also corrected for T1WP. C) and D) are analogous plots for supratentorial total volume. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The use of the percentage change in vCSF (i.e., performing a log-transformation on the data before applying the model) is reasonable since the changes in vCSF are small relative to the vCSF volume and the ventricular volume cannot be negative. An exponential model adjusts for the increasing variability between older subjects. Since sulcal CSF is difficult to estimate, as it requires consistent and accurate "skull stripping," we do not consider it when using vCSF. However, this omission of sulcal CSF could lead to incorrect estimation of the atrophy rate.

The main limitation of statistical modeling with regression in our study is that there is no way to prove that the results obtained represent the "ground truth." Important effects may not be accounted for; for example, although the purpose of our study was not to measure a difference in atrophy rate with different therapeutic modalities, therapy does affect atrophy rate (Borges et al., 2013). A long-term follow-up study taking place on identical hardware with identical protocols and analyzed with the same method presented herein would be the ideal way to verify that our method generates atrophy rates similar to those measured by established methods, but such data are rare and were not available to us. Indeed, as the difficulty in conducting such a study is the problem this research is designed to circumvent, it is easy to see the challenge in comparing these results to a "gold-standard" approach. Additionally, with variability in the methods used to estimate brain tissue volume as well as the natural variability in atrophy between MS patients, it is unclear what meaning can be ascribed to minor differences in the estimation of atrophy rate seen between this and prior atrophy studies.

Short-term or highly variable data, such as the sGM and sWM data from Cohort 1, lead to less accurate subject-specific estimates, regardless of the model being used. In this case, the mixed-effects model estimates atrophy in a similar fashion to a population-level regression. Although this is a technically valid way to generate a population estimate of atrophy rate using these data, the variability present makes the utility of the calculated subject-specific rates unclear. Our model does perform better with longer follow-up length, as can be seen by comparing Cohorts 1 and 2. One prior study (Hughes et al., 2012) found that strong within-cohort rank stability in the EDSS score was

Table 4
Estimates of atrophy rates, with associated 95% confidence intervals (CI) as well as standard deviations (SD) for subject-specific slopes, using the mixed-effects model with and without T1WP correction for ventricular cerebrospinal fluid (vCSF) and total supratentorial (sTot) volumes in Cohort 2. The residual variance of the model is also reported. T1WP, T1-weighted protocol.

Model	vCSF rate [95% CI]	SD vCSF rates	Residual variance [$(\log\text{-mm}^3)^2$]	sTot rate [95% CI]	SD sTot rates	Residual variance
T1WP correction	4.4%/year [3.3, 5.4]	±4.0%/year	2.9×10^{-3}	−4.8 ml/year [−6.5, −3.2]	±6.5 ml/year	540 ml^2
No T1WP correction	4.0%/year [3.0, 5.0]	±3.8%/year	3.1×10^{-3}	−3.5 ml/year [−5.1, −1.9]	±6.2 ml/year	660 ml^2

Table 5
Sample size (SS) necessary in each arm of a therapeutic trial to detect differences of 25%, 50%, and 75% in the atrophy rate using ventricular CSF (vCSF) volume and supratentorial gray matter (sGM) volume, based on the variance of the slopes and residuals from the mixed-effects model fit to Cohort 1 (see text for formula).

Trial length	% diff. in atrophy rate	vCSF SS	sGM SS
1 year	25	160	2828
	50	40	707
	75	18	315
2 years	25	65	712
	50	17	178
	75	8	80

only seen after 4 years of follow-up in a population of all MS subtypes, illustrating that this problem is not limited to our model alone and arises from natural variability in the disease and in the methods of assessing it. At the same time, the need for longer follow-up length competes with other considerations when designing and conducting clinical trials of MS therapeutics that use brain atrophy as a surrogate outcome.

Sample-size calculations using our data support the importance of follow-up length. Specifically, we found that, after adjusting for heterogeneously acquired MRI data, increasing follow-up length provides a decreased sample size requirement for MS clinical trials. These sample sizes are similar to or even improved over other published reports (Altmann et al., 2009; Anderson et al., 2007). As discussed previously, however, high within-subject variance (resulting in high residuals in the formula used to calculate sample size) necessitates a very large sample size and makes the use of direct estimation of the gray matter volume impractical with our method. It is possible that more homogeneously acquired MRI data, coupled with improved cross-sectional and longitudinal (Dwyer et al., 2013) segmentation methods, could greatly decrease the sample size requirement for sGM, but, as discussed, we suggest that vCSF volume is a practical surrogate measure for sGM. In addition to the fact that the data used for the sample-size calculation presented here are not likely to be replicated in any real clinical trial, a realistic trial would also have more than one time point. Unfortunately, for the random intercept/random slope mixed-effects model presented here, there is no known formula to analytically calculate the sample sizes, necessitating numerical simulations that are beyond the scope of this paper. Briefly, with a model similar to ours as the starting point, a fixed treatment effect would be added and levels of this effect on atrophy rate chosen for study. Then, simulated data sets for the time points chosen for the trial would be created with increasing sample size; the sample size at which the power to reject the null hypothesis exceeded a predetermined level would be the sample size necessary for the trial.

One reason that uniform MRI acquisition protocols are generally preferred is that current segmentation techniques tend to perform poorly when tissue contrast changes, even slightly, as a result of variation in acquisition parameters. Although lesion-TOADS allows a direct estimation of brain volume from cross-sectional data, the between-scan variability in the estimation of the gray-white border that arises when acquisition parameters change is high enough to result in poor subject-specific model fits, especially when the data are not collected over a sufficiently long follow-up period. Although we apply a statistical control for protocol type in our model, very small changes in contrast that are not reliably related to a particular protocol, or that are not accounted for in our T1WP code, can still have a large effect on the segmentation results. This property of the segmentation, combined with much higher tissue contrast between ventricles and brain than between gray and white matter, accounts for the more reliable estimation of vCSF volume than sGM or sWM volumes. Our use of average lesion volume rather than lesion volumes at individual time points arises from the same issue, and is limiting in that patients with longer disease durations are likely to have more lesions. This problem highlights the need for longitudinal segmentation techniques that consider all the information available for a given subject when segmenting tissue types; one such method was recently presented (Dwyer et al., 2013). A "smarter" segmentation would recognize, for example, that an outlier result from a particular scan that shows excessive change in gray matter volume compared to prior scans is probably spurious. Such a method might increase the ability of multiple scans in a shorter period of time to detect treatment effects or predict long-term outcomes, possibly lessening the necessary follow-up length and increasing the speed at which new, effective therapies can reach the patients who need them.

5. Conclusions

Linear, multivariable mixed-effects regression can be successfully applied to longitudinal MRI data from multiple scanning protocols in order to generate a coherent estimate of population-level brain volume changes in MS. This method would therefore allow the use of data acquired with non-uniform MRI acquisition parameters, potentially increasing sample size and follow-up length in future MS trials using atrophy as an endpoint and possibly allowing multicenter trials to overcome the obstacle of differing hardware at different sites. Ventricular CSF volume is the most reliably measured brain volume, and our results support other studies that report that gray matter is primarily responsible for atrophy, making the ventricles a more easily quantified surrogate to measure gray matter loss. We believe that mixed-effects modeling of absolute brain volumes can allow a wider use of available data in the study of brain atrophy in MS.

Acknowledgments

The authors thank Henry McFarland and Bibiana Bielekova, who led the studies from which we drew many of the subjects included in this analysis. Subject follow-up and data acquisition could not have occurred without the clinicians, nurses, and staff of the Neuroimmunology Clinic; Roger Stone; John Ostuni; and the NIH FMRIF and Radiology and Imaging Sciences Department staff. BCJ was supported by the Howard Hughes Medical Institute—National Institutes of Health Research Scholars Program. The NINDS Intramural Research Program supported this research.

References

Altmann, D.R., Jasperse, B., Barkhof, F., Beckmann, K., Filippi, M., Kappos, L.D., et al., 2009. Sample sizes for brain atrophy outcomes in trials for secondary progressive multiple sclerosis. Neurology 72, 595–601.

Anderson, V.M., Bartlett, J.W., Fox, N.C., Fisniku, L., Miller, D.H., 2007. Detecting treatment effects on brain atrophy in relapsing remitting multiple sclerosis: sample size estimates. Journal of Neurology 254, 1588–1594.

Antulov, R., Carone, D.A., Bruce, J., Yella, V., Dwyer, M.G., Tjoa, C.W., et al., 2011. Regionally distinct white matter lesions do not contribute to regional gray matter atrophy in patients with multiple sclerosis. Journal of Neuroimaging 21, 210–218.

Bates, D., Maechler, M., Bolker, B., 2011. lme4: Linear Mixed-effects Models Using S4 Classes.

Borges, I., Shea, C., Ohayon, J., Jones, B., Stone, R., Ostuni, J., et al., 2013. The effect of daclizumab on brain atrophy in relapsing–remitting MS. Multiple Sclerosis and Related Disorders 2, 133–140.

Carass, A., Wheeler, M.B., Cuzzocreo, J.L., Bazin, P.-L., Bassett, S.S., Prince, J.L., 2007. A joint registration and segmentation approach to skull stripping. Proceedings of the IEEE International Symposium on Biomedical Imaging 656–659.

Chard, D., Parker, G., Griffin, C., Thompson, A., Miller, D., 2002. The reproducibility and sensitivity of brain tissue volume measurements derived from an SPM-based segmentation methodology. Journal of Magnetic Resonance Imaging 15, 259–267.

Cnaan, A., Laird, N.M., Slasor, P., 1997. Tutorial in biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Statistics in Medicine 16, 2349–2380.

De Stefano, N., Giorgio, A., Battaglini, M., Rovaris, M., Sormani, M.P., Barkhof, F., et al., 2010. Assessing brain atrophy rates in a large population of untreated multiple sclerosis subtypes. Neurology 74, 1868–1876.

Diggle, P., Zeger, S.L., Liang, K.-Y., Heagerty, P., 2002. Analysis of longitudinal data. In: Atkinson, A.C., Copas, J.B., Pierce, D.A., Schervish, M.J., Carroll, R.J., Hand, D.J., et al. (Eds.), Oxford Statistical Science Series. Oxford University Press, p. 379.

Dwyer, M., Bergsland, N., Zivadinov, R., 2013. Improved longitudinal gray matter atrophy assessment via a combination of SIENA and a 4-dimensional hidden Markov random field model. Proc. ISMRM, vol. 21, p. 3014 (Salt Lake City, UT).
Filippi, M., Paty, D.W., Kappos, L., Barkhof, F., Compston, D.A., Thompson, A.J., et al., 1995. Correlations between changes in disability and T2-weighted brain MRI activity in multiple sclerosis: a follow-up study. Neurology 45, 255–260.
Filippi, M., Rovaris, M., Inglese, M., Barkhof, F., De Stefano, N., Smith, S., et al., 2004. Interferon beta-1a for brain tissue loss in patients at presentation with syndromes suggestive of multiple sclerosis: a randomised, double-blind, placebo-controlled trial. Lancet 364, 1489–1496.
Fisher, E., Lee, J.-C., Nakamura, K., Rudick, R.A., 2008. Gray matter atrophy in multiple sclerosis: a longitudinal study. Annals of Neurology 64, 255–265.
Fisniku, L.K., Chard, D.T., Jackson, J.S., Anderson, V.M., Altmann, D.R., Miszkiel, K.A., et al., 2008. Gray matter atrophy is related to long-term disability in multiple sclerosis. Annals of Neurology 64, 247–254.
Fitzmaurice, G., Laird, N., Ware, J., 2011. Applied Longitudinal Analysis, 2nd ed. John Wiley & Sons, Ltd., Hoboken, NJ.
Fox, R.J., Fisher, E., Tkach, J., Lee, J.-C., Cohen, J.A., Rudick, R.A., 2005. Brain atrophy and magnetization transfer ratio following methylprednisolone in multiple sclerosis: short-term changes and long-term implications. Multiple Sclerosis 11, 140–145.
Furby, J., Hayton, T., Altmann, D., Brenner, R., Chataway, J., Smith, K.J., et al., 2010. A longitudinal study of MRI-detected atrophy in secondary progressive multiple sclerosis. Journal of Neurology 257, 1508–1516.
Giorgio, A., Battaglini, M., Smith, S.M., De Stefano, N., 2008. Brain atrophy assessment in multiple sclerosis: importance and limitations. Neuroimaging clinics of North America 18, 675–686.
Horakova, D., Dwyer, M.G., Havrdova, E., Cox, J.L., Dolezal, O., Bergsland, N., et al., 2009. Gray matter atrophy and disability progression in patients with early relapsing–remitting multiple sclerosis: a 5-year longitudinal study. Journal of Neurological Sciences 282, 112–119.
Hughes, S., Spelman, T., Trojano, M., Lugaresi, A., Izquierdo, G., Grand'maison, F., et al., 2012. The Kurtzke EDSS rank stability increases 4 years after the onset of multiple sclerosis: results from the MSBase registry. Journal of Neurology, Neurosurgery & Psychiatry 83, 305–310.
Kappos, L., Moeri, D., Radue, E.W., Schoetzau, A., Schweikert, K., Barkhof, F., et al., 1999. Predictive value of gadolinium-enhanced magnetic resonance imaging for relapse rate and changes in disability or impairment in multiple sclerosis: a meta-analysis. Lancet 353, 964–969.
Martola, J., Bergström, J., Fredrikson, S., Stawiarz, L., Hillert, J., Zhang, Y., et al., 2010. A longitudinal observational study of brain atrophy rate reflecting four decades of multiple sclerosis: a comparison of serial 1D, 2D, and volumetric measurements from MRI images. Neuroradiology 52, 109–117.
Mellanby, A., Reveley, M., 1982. Effects of acute dehydration on computerised tomographic assessment of cerebral density and ventricular volume. Lancet 2, 874.
Miller, D.H., Soon, D., Fernando, K.T., MacManus, D.G., Barker, G.J., Yousry, T.A., et al., 2007. MRI outcomes in a placebo-controlled trial of natalizumab in relapsing MS. Neurology 68, 1390–1401.
Molyneux, P., Kappos, L., Polman, C., Pozzilli, C., Barkhof, F., Filippi, M., et al., 2000. The effect of interferon beta-1b treatment on MRI measures of cerebral atrophy in secondary progressive multiple sclerosis. Brain 123, 2256–2263.
Pirko, I., Lucchinetti, C., Sriram, S., Bakshi, R., 2007. Gray matter involvement in multiple sclerosis. Neurology 68, 634–642.
Polman, C., Reingold, S., Edan, G., Filippi, M., Hartung, H., Kappos, L., et al., 2005. Diagnostic criteria for multiple sclerosis: 2005 revisions to the McDonald criteria. Annals of Neurology 58, 840–846.
Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., et al., 2011. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Annals of Neurology 69, 292–302.
R Development Core Team, 2011. R: A Language and Environment for Statistical Computing.
Rao, A., Richert, N., Howard, T., Lewis, B., Bash, C., McFarland, H., et al., 2002. Methylprednisolone effect on brain volume and enhancing lesions in MS before and during IFNβ-1b. Neurology 59, 688–694.
Roosendaal, S.D., Bendfeldt, K., Vrenken, H., Polman, C.H., Borgwardt, S., Radue, E.W., et al., 2011. Grey matter volume in a large cohort of MS patients: relation to MRI parameters and disability. Multiple Sclerosis 17, 1098–1106.
Rovaris, M., Comi, G., Rocca, M.A., Wolinsky, J.S., Filippi, M., 2001. Short-term brain volume change in relapsing–remitting multiple sclerosis: effect of glatiramer acetate and implications. Brain 124, 1803–1812.
Sampat, M.P., Healy, B.C., Meier, D.S., Dell'Oglio, E., Liguori, M., Guttmann, C.R.G., 2010. Disease modeling in multiple sclerosis: assessment and quantification of sources of variability in brain parenchymal fraction measurements. NeuroImage 52, 1367–1373.
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, 1524–1535.
Shiee, N., Bazin, P.-L., Zackowski, K.M., Farrell, S.K., Harrison, D.M., Newsome, S.D., et al., 2012. Revisiting brain atrophy and its relationship to disability in multiple sclerosis. PLoS One 7, e37049.
Shinohara, R.T., Crainiceanu, C.M., Caffo, B.S., Gaitán, M.I., Reich, D.S., 2011. Population-wide principal component-based quantification of blood–brain-barrier dynamics in multiple sclerosis. NeuroImage 57, 1430–1446.
Shirani, A., Zhao, Y., Kingwell, E., Rieckmann, P., Tremlett, H., 2012. Temporal trends of disability progression in multiple sclerosis: findings from British Columbia, Canada (1975–2009). Multiple Sclerosis 18, 442–450.
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, 87–97.
Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., et al., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17, 479–489.
Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer, New York.
Zivadinov, R., Reder, A.T., Filippi, M., Minagar, A., Stüve, O., Lassmann, H., et al., 2008. Mechanisms of action of disease-modifying agents and brain volume changes in multiple sclerosis. Neurology 71, 136–144.