NeuroImage: Clinical 8 (2015) 606–610

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Handling changes in MRI acquisition parameters in modeling whole brain lesion volume and atrophy data in multiple sclerosis subjects: Comparison of linear mixed-effect models

Alicia S. ChuaMS[a], Svetlana EgorovaMD, PhD[a,b], Mark C. AndersonMS[a], Mariann Polgar-TurcsanyiMS[a], Tanuja ChitnisMD[a,b], Howard L. WeinerMD[a,b], Charles R.G. GuttmannMD[a,d], Rohit BakshiMD[a,b,d], Brian C. HealyPhD[a,b,c,*]

[a] Partners Multiple Sclerosis Center, Brigham and Women's Hospital, Boston, MA, USA
[b] Department of Neurology, Harvard Medical School, Boston, MA, USA
[c] Biostatistics Center, Massachusetts General Hospital, Boston, MA, USA
[d] Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA

ARTICLE INFO

Article history:
Received 29 April 2015
Received in revised form 24 June 2015
Accepted 28 June 2015
Available online 2 July 2015

Keywords:
Multiple sclerosis
MRI
Brain atrophy
T2 lesion
Mixed-effect models

ABSTRACT

Magnetic resonance imaging (MRI) of the brain provides important outcome measures in the longitudinal evaluation of disease activity and progression in MS subjects. Two common measures derived from brain MRI scans are the brain parenchymal fraction (BPF) and T2 hyperintense lesion volume (T2LV), and these measures are routinely assessed longitudinally in clinical trials and observational studies. When measuring each outcome longitudinally, observed changes may be potentially confounded by variability in MRI acquisition parameters between scans. In order to accurately model longitudinal change, the acquisition parameters should thus be considered in statistical models. In this paper, several models for including protocol as well as individual MRI acquisition parameters in linear mixed models were compared using a large dataset of 3453 longitudinal MRI scans from 1341 subjects enrolled in the CLIMB study, and model fit indices were compared across the models. The model that best explained the variance in BPF data was a random intercept and random slope with protocol specific residual variance along with the following fixed-effects: baseline age, baseline disease duration, protocol and study time. The model that best explained the variance in T2LV was a random intercept and random slope along with the following fixed-effects: baseline age, baseline disease duration, protocol and study time. In light of these findings, future studies pertaining to BPF and T2LV outcomes should carefully account for the protocol factors within longitudinal models to ensure that the disease trajectory of MS subjects can be assessed more accurately.

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) that results in impairment of a range of functions, including physical disability and cognitive dysfunction (Krupp et al., 1989; Rao et al., 1991; Steinman, 2001). Magnetic resonance imaging (MRI) of the brain is an established tool to monitor disease activity and disease progression by measuring brain volume loss and lesion load accrual (Miller et al., 1998). Whole brain volume loss, which can be estimated by the brain parenchymal fraction (BPF), is a common measure of neurodegeneration; while, brain lesion load, assessed by the T2 hyperintense lesion volume (T2LV), is a common measure of the total cerebral burden of inflammatory/demyelinating foci in MS (Bermel and Bakshi, 2006; Filippi et al., 2002; Fisher et al., 2002; Wei et al., 2004). Many cross-sectional studies have compared these measures in groups of MS subjects to show that increased T2LV and lower BPF reflect more advanced disease. In addition to cross-sectional studies, many longitudinal studies including the most recent clinical trials have assessed the changes in these measures over time (Rudick et al., 2000; Zivadinov et al., 2007). When measuring cross-sectional and longitudinal change in each MRI measure, researchers must consider the potential impact of between-subject and within-subject variations due to MRI acquisition protocols. To address the potential confounding effects associated with changes in acquisition parameters, most clinical trials require that all of the sites use the same protocols, and significant effort is expended to ensure similarity across scanners.

Even though a well-controlled clinical trial provides the best evidence regarding short-term changes in MRI measures, these studies are generally limited to 2–3 years. To study longer timelines, investigators may need to rely on "real world" observational studies. Such

* Corresponding author at: Partners Multiple Sclerosis Center, 1 Brookline Place, Suite 602, Brookline, MA 02445, USA.
E-mail address: bchealy@partners.org (B.C. Healy).

http://dx.doi.org/10.1016/j.nicl.2015.06.009

observational studies typically involve variations in scanner platform and acquisition protocol to meet the demands of routine clinical care and ever-changing acquisition platforms and protocols. Therefore, there is an unmet need to consider what statistical approaches may be necessary to address the challenge of providing unbiased estimates of intra-subject and inter-subject changes with time in the face of heterogeneously obtained MRI data.

Linear mixed-effect (LME) models (Fitzmaurice et al., 2012; Verbeke and Molenberghs, 2009) have been shown to be efficient in performing group-based inference for neuroimaging data, as shown in studies analyzing both functional MRI (fMRI) and structural MRI data (Bernal-Rusiel et al., 2013; Bowman, 2014; Lange, 1999). LME models account for both between-subject and within-subject variance components, enabling researchers to obtain subject-specific estimated means and account for unbalanced data due to measurements at irregular time points of observation (Fitzmaurice et al., 2012; Verbeke and Molenberghs, 2009). Within the past decade, several MS researchers have aimed to study subject-specific atrophy rates via the mixed-effect modeling framework (Anderson et al., 2007; Jones et al., 2013; Liguori et al., 2011). More recently, Jones et al. demonstrated that a mixed-effect model is superior to a linear regression model in explaining the brain atrophy rate. These investigators proposed including acquisition protocol in LME models using a categorical variable for protocol (Jones et al., 2013). Despite this initial investigation of the impact of protocol on modeling disease course in MS, a comprehensive evaluation of the possible fixed and random effects including those associated with acquisition parameters has not been completed. Such assessment is warranted to appropriately identify the mean response trajectory for MRI data such as BPF and T2LV. The aim of this study was to build a comprehensive model for the BPF and T2LV including subject-specific (random) effects and MRI acquisition parameter (individually or combined) (fixed) effects.

2. Methods

2.1. Subjects

Longitudinal MRI scans from 1341 subjects were selected from the Comprehensive Longitudinal Investigation of Multiple Sclerosis at the Brigham and Women's Hospital, Partners MS Center (CLIMB), an ongoing prospective observational cohort study that began enrolling subjects in 2000 (Gauthier et al., 2006). Inclusion criteria for the CLIMB study are age ≥18 years and a clinically isolated syndrome (CIS) or diagnosis of MS according to the revised McDonald criteria (Polman et al., 2005). Subjects have clinical visits every 6 months that include complete neurological examinations and Expanded Disability Status Scale (EDSS) ratings (Kurtzke, 1983). All subjects from the CLIMB cohort with available MRI scans up to 6 years after study entry were included in this study. We limited the follow-up time to 6 years in order to ensure that the subset of subjects followed for much longer periods did not have too much leverage on this analysis. Hence, scans for subjects that are greater than 6 years from their baseline scan were excluded. Demographic and clinical characteristics of subjects are provided in Table 1.

2.2. Image acquisition and processing

BPF and T2LV were calculated by applying an image analysis pipeline to dual-echo, conventional spin-echo images (DE-CSE). Over the years, some acquisition parameter variations were included in the DE-CSE pulse sequence. All MRI scans were acquired on various 1.5 T Signa GE scanners at the Brigham and Women's Hospital (BWH), Boston, Massachusetts, using at times a standard quadrature head coil, and at other times a multichannel head coil (8 Channel High Resolution Brain Array (8HR BRAIN)). DE-CSE MRI protocols were acquired axially with pulse sequences including various combinations of following parameter ranges: TR = 2216–3000 ms, TE1/TE2 30/80 ms, slice thickness 3 mm, with no interslice gaps, resulting in a pixel size of 0.7813–0.9375 mm. In this study, nineteen unique DE-CSE MRI protocols (i.e., variably parametrized DE-CSE MRI pulse sequences) were used as detailed in Table 2. The original standard protocol in the CLIMB study was protocol A. Over the years, this protocol had to be adapted because of operational considerations, resulting in the 19 distinct protocols in Table 2. Quantitative image analysis was performed from the dual-echo images using an automated template-driven segmentation pipeline with partial volume effect correction (TDS+) (Wei et al., 2002) followed by manual editing of output segmentation maps by an experienced observer. Large scale data management and analysis was enabled by an image analysis workflow management system linked to our Image Centered Oracle MS Database (Liu et al., 2005). The pipeline involves a semi-automated skull-stripping editing procedure to derive the intracranial volume (ICV) followed by automated segmentation of gray matter (GM), CSF, white matter (WM), and white matter lesions through TDS+ (Wei et al., 2002). BPF was calculated by the following formula: BPF = (GM + WM + lesions)/ICC (Wei et al., 2004), where ICC is the volume of the intracranial cavity serving as reference for individual head size (Kikinis et al., 1992). After the pipeline was completed, all of the MRI scans reported in this study (N = 3453) underwent manual correction of automatically generated segmentation maps of T2 lesions, brain parenchymal compartments and CSF by expert readers using 3D Slicer software (Liu et al., 2005). We note that no correction for misclassification of T1 hypointensities was performed in our pipeline, but recent work from our group has demonstrated that these hypointensities have a limited impact on measures of BPF (Dell'Oglio et al., 2015).

Table 1
Demographic characteristics of study subjects.

N subjects		1341
N scans		3453
% males		26.0
Baseline visit age (years, mean(SD))		43.7 (11.2)
Race (%)		
	American Indian/Alaska Native	0.2
	Asian	0.5
	Black/African-American	3.0
	More than one race	1.8
	Native Hawaiian/Pacific Islander	0.1
	White	92.8
	Unknown/unreported	1.6
Ethnicity (%)		
	Hispanic or Latino	3.4
	Non-Hispanic or Latino	95.1
	Unknown/unreported	1.5
Disease category (%)		
	RRMS	73.3
	PPMS	5.1
	SPMS	17.2
	PRMS	1.0
	CIS	3.3
EDSS closest to baseline scan (median, IQR)		1.5 (0.0, 2.5)
Median scans per person (n, range)		2.0 (1.0–12.0)
Mean follow-up time (years, mean(SD))		1.6 (1.8)

Legend: RRMS: relapsing–remitting multiple sclerosis; PPMS: primary progressive multiple sclerosis; SPMS: secondary progressive multiple sclerosis; PRMS: progressive relapsing multiple sclerosis; CIS: clinically isolated syndrome; EDSS: expanded disability status scale.

2.3. Statistical analysis

Statistical analyses were conducted using the MIXED procedure in the Statistical Analysis System (SAS) version 9.3 (Cary, NC). A cube

Table 2
Magnetic resonance imaging acquisition parameters in the CLIMB study.

Protocol	N scans	Coil	Pixel spacing (mm)	Pixel bandwidth (mm)	Repetition time (ms)	Scanner Name
A	2125	Head	0.9375	84.9219	3000	BWH[a]: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
B	41	Head	0.9375	84.9219	3000	BWH: Pike 2 — 75 Francis St., Boston, MA (1.5 T Scanner)
C	210	Head	0.9375	81.4062	3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
D	63	Head	0.9375	84.9219	<3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
E	75	8 HR Brain[b]	0.9375	84.9219	3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
F	31	8 HR Brain	0.9375	81.4062	3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
G	104	8 HR Brain	0.8594	84.9219	3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
H	30	8 HR Brain	0.8594	81.4062	3000	BWH: Pike 1 — 75 Francis St., Boston, MA (1.5 T Scanner)
I	24	8 HR Brain	0.8594	81.4062	3000	BWH: Pike 2 — 75 Francis St., Boston, MA (1.5 T Scanner)
J	181	8 HR Brain	0.8594	84.9219	<3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
K	13	8 HR Brain	0.8594	81.4062	<3000	BWH: Pike 1 — 75 Francis St., Boston, MA (1.5 T Scanner)
L	47	8 HR Brain	0.8594	81.4062	<3000	BWH: Pike 2 — 75 Francis St., Boston, MA (1.5 T Scanner)
M	225	8 HR Brain	0.7813	84.9219	3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
N	89	8 HR Brain	0.7813	81.4062	3000	BWH: Pike 1 — 75 Francis St., Boston, MA (1.5 T Scanner)
O	19	8 HR Brain	0.7813	81.4062	3000	BWH: Pike 2 — 75 Francis St., Boston, MA (1.5 T Scanner)
P	136	8 HR Brain	0.7813	84.9219	<3000	BWH: 221 Longwood Ave., Boston, MA (1.5 T Scanner)
Q	16	8 HR Brain	0.7813	81.4062	<3000	BWH: Pike 1 — 75 Francis St., Boston, MA (1.5 T Scanner)
R	12	8 HR Brain	0.7813	81.4062	<3000	BWH: Pike 2 — 75 Francis St., Boston, MA (1.5 T Scanner)
S	12	8 HR Brain	0.8203	81.4062	3000	BWH: Pike 1 — 75 Francis St., Boston, MA (1.5 T Scanner)

[a] BWH: Brigham and Women's Hospital, Boston, MA.
[b] 8 HR Brain: 8 Channel High Resolution Brain Array.

root transformation was applied to T2LV prior to analysis to avoid violation of model assumptions. In addition, BPF was modeled as a percent to allow easier interpretation of the model coefficients. Since the goal of our analysis was the estimation of both the mean BPF and T2LV trajectories over time with a set of serial measurements from subjects, an LME model was used. In the process of model selection, several models were proposed and investigated, but two main considerations were of interest. First, using the most complex mean model, several potential models for the covariance/random effect structures were compared. Once the covariance/random effects structures were compared, several models for the impact of protocol parameterization on the fixed effects were assessed (Diggle et al., 2002).

For the assessment of the covariance structure, the model included protocol, baseline age, baseline disease duration, study time and an interaction term between protocol and study time. In terms of covariance structure/ random effects, we compared three models: (A1) subject-specific random intercept only, (A2) subject-specific random intercept and study time effect, and (A3) subject-specific intercept and study time effect with protocol specific residual variance. Although assessment of protocol specific random study time effects was of interest, many protocols failed to have multiple observations on the same subject so these variance components were not estimable in our dataset. In order to compare models, Akaike Information Criterion (AIC) with the "smaller-is-better" criterion was used. The regression equation for each model and the associated SAS code for fitting the model are shown in the Appendix.

After selection of the best covariance parameter model, selection of the fixed effects was performed based on five possible models. The first model (Model B1) included age at the time of the visit, baseline disease duration and protocol as fixed effects. Each of the 19 protocols in the study was given a separate intercept in this model. This model is the same as the model from Jones et al. without the additional covariates. Despite the appeal of using age at scan as the time metric, this model makes the assumption that the cross-sectional and longitudinal effects of age are the same. To assess this assumption, study age was broken into two components (baseline age and study time) in Model B2, but protocol remained in the model as a fixed effect. The third model (Model B3) added protocol by study time interactions to the previous model (Model B2), to assess whether the estimated change with time was different across the protocols. The fourth and fifth models (Models B4 and B5) were similar to Models B2 and B3, but rather than using a separate intercept for each of the protocols, the components of the protocols were included as separate fixed effects. This approach reduced the number of parameters to estimate but added the assumption that the effect of each component of the protocol was independent and additive. For the protocols in this study, protocols were defined using these parameters: type of coil, pixel bandwidth, pixel size, repetition time, and scanner (Table 2). We note that echo time was always the same across all protocols so this parameter was not included. The regression equation for each of the models and SAS code for fitting the models are shown in the Appendix. In order to compare the models, the AIC was used, as described above.

3. Results

3.1. BPF

Comparison of models for the selection of covariance parameters is presented in at the top of Table 3; additional statistics from each of the models are presented in Supplementary Table 1. The random intercept and slope model (Model A2) was superior to the random intercept only model (Model A1), and the improvement was substantial. Further, the random intercept and slope model with protocol specific residual variance (Model A3) was superior to the random intercept and slope model. Therefore, Model A3 was the chosen covariance structure.

Using this covariance structure, the fit of the model with the specified fixed effects are shown in Table 3; additional statistics from each of the models are presented in Supplementary Table 2. The results show that the lowest AIC was attained with Model B2. This model includes separate estimates for cross-sectional effect of age and the

Table 3
Model comparisons for brain parenchymal fraction.

Model comparison	Smaller model AIC	Larger model AIC	Preferred model
Covariance models			
A1 vs. A2	14,390.1	14,149.4	A2
A1 vs. A3	14,390.1	13,995.4	A3
A2 vs. A3	14,149.4	13,995.4	A3
Mean models			
B1 vs. B2	14,029.6	13,983.0	B2
B2 vs. B3	13,983.0	13,992.1	B2
B4 vs. B2	14,004.5	13,983.0	B2
B5 vs. B3	14,009.7	13,992.1	B3
B4 vs. B5	14,004.5	14,009.7	B4

Legend: AIC: Akaike Information Criterion; for full description of models A1–A3 and models B1–B5, please refer to the Statistical analysis section.

longitudinal change with age, demonstrating the potential problems associated with using age at the MRI visit as the time metric in longitudinal models for BPF. Further, a protocol by study time interaction was not found to add significantly to the model since Model B3 did not lead to improvement over Model B2. Finally, the model including fixed effects for each of the components of the protocol provided an inferior fit compared to a model with a protocol specific effect, demonstrating that the parameters for the protocols do not have simple additive effects on BPF, as might have been expected, given that individual protocol parameters might have complex interactions towards the resulting image contrast and noise characteristics. The final parameter estimates from model B2 are shown in Supplementary Table 3.

3.2. T2LV

Comparison of models for the selection of covariance parameters is presented at the top of Table 4; additional statistics for each of the fixed effects models are shown in Supplementary Table 4. The random intercept and slope model (Model A2) was superior to the random intercept only model (Model A1), and the improvement was substantial. For T2LV, the random intercept and slope model with protocol specific residual variance (Model A3) could not be fit because the residual variance associated with one of the protocols was estimated to be equal to 0. Given the inability to estimate some of the parameters, Model A2 was chosen for further analysis.

Using this covariance structure, the fit of the model with the specified fixed effects are shown in Table 4; additional statistics for each of the fixed effects models are shown in Supplementary Table 5. The results show that the lowest AIC was attained with Models B2 and B5. As for BPF, the results show that including separate estimates for cross-sectional effect of age and the longitudinal change with age showed superior fit compared to a model using age at the MRI visit as the time metric. Further, a protocol by study time interaction was not found to add significantly to the model since Model B3 did not lead to improvement over Model B2. Finally, the model including fixed effects for each of the components of the protocol provided an inferior fit relative to a model with a protocol specific effect. Interestingly, Model B5 had a similar AIC as Model B2, but Model B2 is easier to interpret so this model is chosen as superior. The final parameter estimates from Model B2 are shown in Supplementary Table 6.

4. Discussion

The aim of this study was to build comprehensive models for the estimation of mean BPF and T2LV in MS subjects encompassing subject-specific (random) effects and acquisition parameter (fixed) effects. A series of models with different covariance parameters and models of different fixed-effects were compared, and the optimal model for BPF and T2LV was the same in terms of the fixed effects but differed in terms of the variance components. For the BPF, the variance components included a random intercept and slope as well as protocol specific variance terms, indicating that the residual variability associated with each of the protocols differed. For the T2LV, the random intercept and slope model with equal variance across the protocols was chosen due to the inability to estimate all the protocol specific variance parameters. For each of the outcomes, protocol by study time interactions were not found to significantly improve the models, but separate parameters for the cross-sectional effect of age at study entry and the within subject longitudinal change provided a superior fit compared to a model with a single parameter for age.

Table 4
Model comparisons for cube root transformed T2 lesion volume.

Model comparison	Smaller model AIC	Larger model AIC	Preferred model
Covariance models			
A1 vs. A2	704.8	475.6	A2
A1 vs. A3	704.8	Failed to converge	A1
A2 vs. A3	475.6	Failed to converge	A2
Mean models			
B1 vs. B2	322.2	292.8	B2
B2 vs. B3	292.8	302.7	B2
B2 vs. B4	292.8	302.9	B2
B3 vs. B5	302.7	290.8	B5
B4 vs. B5	302.9	290.8	B5

Legend: AIC: Akaike Information Criterion; for full description of models A1–A3 and models B1–B5, please refer to the Statistical analysis section.

Within the analysis of the BPF, the residual variability associated with each of the protocols was found to differ, demonstrated by the improvement in model fit comparing Models A2 and A3. This result indicates that the homoscedasticity assumption of many commonly used models may be inefficient when subjects are measured using different scanning protocols. When the estimated residual variances were investigated in Supplementary Table 3, the protocols with the largest deviations from protocol A had the largest difference in terms of residual variability. These results show that accounting for heteroscedasticity due to protocol may be an important consideration in modeling BPF data from multiple protocols. In addition to the impact of protocol on residual variance, protocol was found to have an impact on the intercept ($p < 0.001$ for overall effect of protocol in Model B2), but there was no protocol by study time interaction ($p = 0.06$ from Model B3). Further, including the components of the protocol as additive fixed effects failed to improve model relative to including protocol specific fixed effects.

Within the analysis of the cube root transformation of T2LV, the model with protocol specific residual variance failed to converge because the residual variance for one of the protocols was found to be equal to 0. When this protocol was removed, the model with protocol specific residual variance was observed to lead to improved fit, but this model failed to converge with other fixed effects. Therefore, the model with just a random intercept and slope was chosen as optimal based on our dataset, but heteroscedastic variance might be appropriate in other datasets. In addition to the variance components, the comparison of the fixed effects models showed that protocol had a significant effect ($p < 0.001$ for the overall effect of protocol in Model B2), but there was no protocol by study time interaction ($p = 0.10$ from Model B3). Interestingly, a larger number of protocols had a highly significant effect on the intercept for the T2LV compared to the BPF. This results shows that the impact of protocol appears larger for the T2LV, demonstrating the importance of incorporating this into the models.

For both outcome measures, separate parameters for the cross-sectional effect of age and the within subject change with age were found to lead to an improved fit relative to a single parameter for the effect of age. In both models, the within subject change in the outcomes with age was found to be larger than the cross-sectional effect of age. Therefore, estimating the change with age using a single parameter would underestimate the change with time for a specific subject. Further work assessing the impact of age may provide more insight regarding this finding.

For this analysis, model selection was based on the AIC. The AIC includes a penalty for model complexity, but an alternative measure for model selection is the Bayesian Information Criterion (BIC). The BIC has a larger penalty for model complexity compared to the AIC; therefore, models with fewer parameters are favored by the BIC more than the AIC. The BIC for each of the model compared in this paper are also provided in the supplementary tables. When the BIC is used for model selection, the same variance components were chosen, but Model B4 was found to be superior to Model B2 in each case. This result is driven by the fact that Model B4 required fewer parameters. At the same time, we prefer Model B2 because we had sufficient sample size to estimate all of the parameters in our model so choosing the model with fewer parameters was not viewed as an advantage.

Our study has several limitations that warrant further discussion. First, our results are based on a specific MRI processing pipeline. In particular, the reported results regarding brain atrophy are based on analysis of BPF measured two channel (PD/T2 weighted MRI) pipeline in which two variables (lesion volume and brain atrophy) are measured simultaneously, while other pipelines measure normalized brain volume using alternative approaches. To assess whether these results apply to other brain atrophy measures and processing pipelines, similar analyses must be completed. Second, we were unable to assess the impact of protocol on the random effect variances due to the limited number of subjects who had repeated observations on the many protocols. Furthermore, we did not assess the potential role of changes in the post-processing pipeline. This may be especially relevant to the combining of existing datasets from multiple centers that have already been processed by different analysis pipelines. Thus, future work will be required to fully assess the range of deviations that may have an impact on the random effect distributions, which includes the inter-rater reliability of scan editing for both ICC and final segmentation correction. Third, the presence of steroid treatment for a relapse at the time of the MRI scan could have impacted the modeling of longitudinal change. Therefore, future work investigating the impact of steroids/relapses at the time of an MRI scan in modeling the longitudinal change of BPF and T2LV is warranted. Finally, we note that this is an observational study following subjects longitudinally. Our analysis only assessed the impact of protocol parameters described in Table 2, but other potential sources of variability including scanner changes could have impacted the longitudinal changes. Therefore, the impact of other sources of variability on the modeling of longitudinal changes will be a subject of further research.

In conclusion, we believe that our proposed models for both outcomes provide a good fit to the data. In light of these findings, future research pertaining to BPF and T2LV outcomes should carefully account for protocol in the analysis to ensure that the true disease trajectory of MS subjects can be assessed.

Acknowledgments

The CLIMB study is funded in part by the Nancy Davis Foundation and Merck Serono S.A.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.nicl.2015.06.009.

References

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. J. Neurol. 254 (11), 1588–1594. http://dx.doi.org/10.1007/s00415-007-0599-317940723.

Bermel, R.A., Bakshi, R., 2006. The measurement and clinical relevance of brain atrophy in multiple sclerosis. Lancet Neurol. 5 (2), 158–170. http://dx.doi.org/10.1016/S1474-4422(06)70349-016426992.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., Alzheimer's Disease Neuroimaging Initiative, 2013. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. Neuroimage 66, 249–260. http://dx.doi.org/10.1016/j.neuroimage.2012.10.06523123680.

Bowman, F.D., 2014. Brain imaging analysis. Annual Review of Statistics and its Application 1, 61–85. http://dx.doi.org/10.1146/annurev-statistics-022513-11561125309940.

Dell'Oglio, E., Ceccarelli, A., Glanz, B.I., Healy, B.C., Tauhid, S., Arora, A., Neema, M., 2015. Quantification of global cerebral atrophy in multiple sclerosis from 3T MRI using SPM: the role of misclassification errors. J. Neuroimag. 25 (2), 191–199. http://dx.doi.org/10.1111/jon.1219425523616.

Diggle, P.J., Heagerty, P., Liang, K.-Y., Zeger, S., 2002. Analysis of longitudinal data. Oxford University Press.

Filippi, M., Dousset, V., McFarland, H.F., Miller, D.H., Grossman, R.I., 2002. Role of magnetic resonance imaging in the diagnosis and monitoring of multiple sclerosis: consensus report of the White Matter Study Group. J. Magn. Reson. Imaging 15 (5), 499–50411997889.

Fisher, E., Rudick, R.A., Simon, J.H., Cutter, G., Baier, M., Lee, J.C., Simonian, N.A., 2002. Eight-year follow-up study of brain atrophy in patients with MS. Neurology 59 (9), 1412–1420. http://dx.doi.org/10.1212/01.WNL.0000036271.49066.0612427893.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2012. Applied Longitudinal Analysis 998. John Wiley & Sons.

Gauthier, S.A., Glanz, B.I., Mandel, M., Weiner, H.L., 2006. A model for the comprehensive investigation of a chronic autoimmune disease: the multiple sclerosis CLIMB study. Autoimmun. Rev. 5 (8), 532–536. http://dx.doi.org/10.1016/j.autrev.2006.02.01217027888.

Jones, B.C., Nair, G., Shea, C.D., Crainiceanu, C.M., Cortese, I.C., Reich, D.S., 2013. Quantification of multiple-sclerosis-related brain atrophy in two heterogeneous MRI datasets using mixed-effects modeling. Neuroimage Clin. 3, 171–17924179861.

Kikinis, R., Shenton, M.E., Gerig, G., Martin, J., Anderson, M., Metcalf, D., Lorensen, W., 1992. Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging. J. Magn. Reson. Imaging 2 (6), 619–629. http://dx.doi.org/10.1002/jmri.18800206031446105.

Krupp, L.B., LaRocca, N.G., Muir-Nash, J., Steinberg, A.D., 1989. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch. Neurol. 46 (10), 1121–1123. http://dx.doi.org/10.1001/archneur.1989.005204601150222803071.

Kurtzke, J.F., 1983. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33 (11), 1444–1452. http://dx.doi.org/10.1212/WNL.33.11.14446685237.

Lange, N., 1999. Statistical procedures for functional MRI. Functional M.R.I. 301–335.

Liguori, M., Healy, B.C., Glanz, B.I., Khoury, S.J., Moscufo, N., Weiner, H.L., Guttmann, C.R., 2011. HLA (A—B—C and —DRB1) alleles and brain MRI changes in multiple sclerosis: a longitudinal study. Genes Immun. 12 (3), 183–190. http://dx.doi.org/10.1038/gene.2010.5821179117.

Liu, L., Meier, D., Polgar-Turcsanyi, M., Karkocha, P., Bakshi, R., Guttmann, C.R., 2005. Multiple sclerosis medical image analysis and information management. J. Neuroimag. 15 (4 Suppl), 103S–117S. http://dx.doi.org/10.1177/105122840528286416385023.

Miller, D.H., Grossman, R.I., Reingold, S.C., McFarland, H.F., 1998. The role of magnetic resonance techniques in understanding and managing multiple sclerosis. Brain 121 (1), 3–24. http://dx.doi.org/10.1093/brain/121.1.39549485.

Polman, C.H., Reingold, S.C., Edan, G., Filippi, M., Hartung, H.P., Kappos, L., Wolinsky, J.S., 2005. Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald criteria". Ann. Neurol. 58 (6), 840–846. http://dx.doi.org/10.1002/ana.2070316283615.

Rao, S.M., Leo, G.J., Bernardin, L., Unverzagt, F., 1991. Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction. Neurology 41 (5), 685–691. http://dx.doi.org/10.1212/WNL.41.5.6852027484.

Rudick, R.A., Fisher, E., Lee, J.C., Duda, J.T., Simon, J., 2000. Brain atrophy in relapsing multiple sclerosis: relationship to relapses, EDSS, and treatment with interferon beta-1a. Mult. Scler. 6 (6), 365–37211212130.

Steinman, L., 2001. Multiple sclerosis: a two-stage disease. Nat. Immunol. 2 (9), 762–764. http://dx.doi.org/10.1038/ni0901-76211526378.

Verbeke, G., Molenberghs, G., 2009. Linear Mixed Models for Longitudinal Data. Springer.

Wei, X., Guttmann, C.R., Warfield, S.K., Eliasziw, M., Mitchell, J.R., 2004. Has your patient's multiple sclerosis lesion burden or brain atrophy actually changed? Mult. Scler. 10 (4), 402–406. http://dx.doi.org/10.1191/1352458504ms1061oa15327037.

Wei, X., Warfield, S.K., Zou, K.H., Wu, Y., Li, X., Guimond, A., Guttmann, C.R., 2002. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. J. Magn. Reson. Imaging 15 (2), 203–209. http://dx.doi.org/10.1002/jmri.1005311836778.

Zivadinov, R., Locatelli, L., Cookfair, D., Srinivasaraghavan, B., Bertolotto, A., Ukmar, M., Zorzon, M., 2007. Interferon beta-1a slows progression of brain atrophy in relapsing–remitting multiple sclerosis predominantly by reducing gray matter atrophy. Mult. Scler. 13 (4), 490–501. http://dx.doi.org/10.1177/135245850607044617463072.