Redefining multiple sclerosis phenotypes using MRI

Arman Eshaghi[1,2], Alexandra Young[2], Peter Wijertane[2], Ferran Prados[1,2], Douglas Arnold[3], Sridar Narayanan[3], Charles R. G. Guttmann[4], Frederik Barkhof[1,2,5], Daniel C Alexander[2], Alan J Thompson[1,6], Declan Chard*[1], Olga Ciccarelli*[1], *on behalf of the International Progressive MS Alliance (PMSA) Investigators***

[1] Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, WC1B5EH, UK

[2] Centre for Medical Image Computing (CMIC), Department of Computer Science, Faculty of Engineering Sciences, University College London, UK

[3] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada

[4] Center for Neurological Imaging, Brigham and Women's Hospital, Harvard Medical School, Massachusetts, USA

[5] VU University Medical Centre, Amsterdam, The Netherlands

[6] Department of Brain Repair and Rehabilitation, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, UK

* Joint senior authors

** Listed in the appendix of this article

Abstract

Background

There are 4 courses of multiple sclerosis (MS): clinically-isolated syndrome (CIS), relapsing-remitting MS (RRMS), primary-progressive MS (PPMS) and secondary-progressive MS (SPMS). We aimed to achieve a further sophistication in the definition of MS phenotypes by identifying patient subgroups who accumulate magnetic resonance imaging (MRI) abnormalities with similar patterns. We assessed whether data-driven subtyping predicted clinical outcome and response to experimental treatments.

Methods

In this retrospective study, we included longitudinal data from 8,545 people with MS who had 31,451 visits from 14 double-blind randomised controlled trials and three observational cohorts. We included cross-sectional data from 14,928 healthy volunteers. For each visit, we processed brain MRI scans. We obtained 18 MRI variables, includ that included the volume of the cortex of each lobe and deep grey matter, cerebellar grey matter and wite matter, total lesion volume, cerebral white matter, brainstem, and T1/T2 ratio in regions of the normal appearing white matter (or NAWM). We trained a machine learning algorithm, called SuStaIn, on 14 datasets to identify data-driven subtypes. We then tested it in three external, independent datasets.we assessed the external validity of our model by testing if we could predict 24-week confirmed Expanded Disability Status Scale (EDSS) progression, disease activity, and the reduction in EDSS worsening in each subtype on treatment *vs* placebo. To investigate whether there was a difference in the treatment response between the SuStaIn subtypes we used linear mixed effect modles

Findings

We identified three data-driven subtypes with a distinct neuroanatomical spread of abnormality and termed subtypes according to areas they affected early in disease course: cortex-first (44% of patients), NAWM-first (30%), and the lesion-first (26%). Data-driven subtyping and staging, but not clinical classifications or EDSS at baseline, was associated with time to EDSS progression ($\beta_{Subtype=}$0.04 and $\beta_{stage=}$-0.06, p<0.01 for all). In external datasets, compared to

the cortex-first, the lesion-first subtype had a shorter time to EDSS progression (hazard ratios range=1.75 to 1.82 across trials, $p<0.01$) and higher disease activity ($p<0.05$ for relapse rate and gad-enhancing lesions). In three progressive MS trials the lesion-first subtype showed the greatest treatment response (64% more than the cortex-first, $p<0.001$).

Interpretation

Data-driven MS subtypes and stages, when compared with clinical classification or baseline EDSS better predicts the subsequent clinical course and treatment response. Data-driven subtyping has the potential to prospectively enrich clinical trials with patients who are more likely to respond to treatments.

Funding

International Progressive MS Alliance.

Research in Context

Evidence before this study

We searched the Medline database for entries with “multiple sclerosis” and “clinical classification” keywords in their title or abstract published before February 2019. We found two consensus-based recommendations defining MS disease courses. First, in 1996, Lublin and colleagues who devised clinical classification for multiple sclerosis (MS) based on clinical evolution into relapse-onset or progressive onset course. This definition was revised in 2013 with specific terms for relapsing-remitting, secondary progressive and primary progressive terms while considering underlying processes from clinical or brain imaging assessments (progression and disease activity). However, the clinical classification of MS is retrospective and the boundary between MS phenotypes remains subjective mainly based on clinical evolution.

Added value of this study

Here we applied machine learning methods to brain magnetic-resonance imaging to identify data-driven subgroups of patients who share similar patterns in accumulation of abnormalities in large cohorts from recent double-blind randomised controlled trials of MS. We found three subtypes with a distinct neuroanatomical spread of abnormality. These data-driven subtypes could predict disability progression, disease activity, and treatment response in relapse-onset and progressive MS and even within patients with a single clinical subtype.

Implications of the all available evidence

Data-driven MS subtyping has the potential to enrich trial cohorts with those who have the most potential to demonstrate treatment effects.

Introduction

There are four courses of multiple sclerosis (MS): clinically-isolated syndrome (CIS), relapsing-remitting MS (RRMS), primary-progressive MS (PPMS) and secondary progressive MS (SPMS).[1] These clinical subtypes are based on the current patient status and historical data. In 2013 a revised MS phenotype description has added two modifiers to the basic MS phenotypes: disease activity (relapses and/or magnetic resonance imaging (MRI)) and progression of disability[2]. Since these additional descriptors reflect acute inflammation and neurodegeneration, they may help to prognosticate patient outcomes and enhance homogeneity of disease course in clinical trials. The four clinical phenotypes of MS and their additional modifiers are routinely used in the clinical setting and clinical trials.

When examining the clinical, imaging, immunologic, or pathologic characteristics of the four clinical courses of MS, it is clear that the differences between the MS subtypes are relative rather than absolute. SPMS and PPMS share more similarities than differences in MRI features and pathogenic mechanisms.[3] A CIS phenotype may evolve into RRMS, and RRMS phenotype may transition into SPMS.[2,4] Therefore, different phenotypes are often used together in clinical trials[5,6]. When these transitions happen exactly is challenging to ascertain, because they are often based on the *subjective* recollection of symptoms and their evolution. Subgroup analyses of clinical trials have often detected treatment effects not seen when all patients were analysed together[6,7]. Overall, these considerations suggest the need for a greater sophistication in the definition of MS phenotypes.

New opportunities for further progress in defining the phenotypes of MS may arise if *objective* indicators of a patient's biological status are used. MRI features are closer to the biology of MS than clinical symptoms, and MRI abnormalities are the manifestations of pathogenetic mechanisms occurring in MS.[8] The overarching hypothesis of our work is that the definition of MS and its subtypes can be linked with the biology of the disease, as reflected by MRI abnormalities, instead of clinical features. The consequence of this new framework, which

does not define MS phenotypes on the basis of clinical symptoms and course of disability over time, but describes them as a consequence of the pathological changes visible on MRI scans, will represent an important step towards personalised medicine, since it will lead to the use of therapies to target patient populations who share the same pathogenic mechanisms of the disease.[9]

To redefine the MS phenotypes on the basis of MRI abnormalities we analysed a large number of MRI scans in patients with RRMS, SPMS and PPMS, using a machine learning algorithm, called Subtype and Staging Inference (SuStaIn),[10] which can disentangle the different patterns of MRI changes. We have recently developed and tested this algorithm in patients with Alzheimer's disease.[10] SuStaIn identifies data-driven patient subtypes who accumulate MRI abnormalities with similar patterns and enables identification of subgroups using cross-sectional MRI data. SuStaIn determines how closely information from a given patient matches 'learned' data-driven subtypes, and what stage the given patient has reached at a particular time.

Here we performed MRI-driven subtyping of patients by pooling data from several randomised controlled trials in MS and tested the external generalisability of our model in an independent data set using SuStaIn. We aimed to: (1) Define MS subtypes on the basis of MRI-derived patterns and test whether they were associated with clinical outcome more strongly than standard MS phenotypes; (2) Test whether there was a difference in treatment response between MRI-driven subtypes; and (3) Stage patients along the trajectory of MRI changes to predict clinical outcomes.

Material and methods

Participants

We carried out this retrospective study under the auspices of the International Progressive MS Alliance (www.progressivemsalliance.org). We collected clinical and magnetic resonance imaging (MRI) data from the following 14 MS randomised-controlled trials (RCTs): five trials of PPMS[7,11–13], five trials of SPMS[14–18], and four trials of RRMS[19–21] (**Table 1**). We also gathered three observational cohorts with mixed MS subtypes,[7,11–26] and two datasets from healthy volunteers.

Each RCT and observational study had received ethical approval and patients had given written, informed consents at the time of data-acquisition. The Institutional Review Board at the Montreal Neurological Institute (MNI), Quebec, Canada approved this study (Reference number: IRB00010120). The pharmaceutical companies that provided the fully anonymised, individual patient raw data, signed an agreement with the McGill University, which allowed pooling data but not testing treatment response in individual RCT.

Healthy volunteers

We downloaded unprocessed MRI data from: (1) The S1200 Open Access release of the Human Connectome Project, (2) The UK Biobank data, which were available for download on 1st of February 2019; this project was approved by the UK Biobank (Reference number: 47233).

Table 1. Datasets used in this study.

Study name	Population	Design	Number of participants with eligible MRI data	Total visits with MRI	Published protocol citation number
Datasets to develop norms					
UK Biobank*	Healthy volunteers	Observational	13,823	13,823	[22]
Human Connectome Project*	Healthy volunteers	Observational	1,105	1,105	[23]
MS datasets in train and cross-validation set**					
Siena	Mixed	Observational	149	595	[41]
Basel	Mixed	Observational	81	239	[25]
DEFINE-CONFIRM, ENDORSE	RRMS	RCT (phase III)	1,071	5,208	[19]
OPERA 1	RRMS	RCT (phase III)	801	3,025	[20]
OPERA 2	RRMS	RCT (phase III)	824	3,044	[20]
ASCEND	SPMS	RCT (phase III)	1,002	5,095	[42]
Lipoic acid	SPMS	RCT (phase II)	41	111	[17]
MS-STAT	SPMS	RCT (phase II)	131	373	[15]

MAESTRO 3	SPMS	RCT (phase III)	539	1,753	[18]
Lamotrigine	SPMS	RCT (phase II)	97	251	[16]
ARPEGGIO	PPMS	RCT (phase II)	409	946	Clinialtrials.gov ID: NCT02284568
INFORMS	PPMS	RCT (phase III)	323	758	[12]
PROMISE	PPMS	RCT (phase III)	458	740	[11]
OLYMPUS	PPMS	RCT (phase II/III)	396	1,630	[7]
MS datasets in the external test set**					
CLIMB	Mixed	Observational	319	1,950	[26]
ORATORIO	PPMS	RCT (phase III)	701	2,724	[13]
BRAVO	RRMS	RCT (phase III)	1,203	3,009	[21]

Table Caption:

***** UK Biobank and Human Connectome Project are cross-sectional cohorts. All others are longitudinal.

** We chose training and external validation sets *a priori*.

*** Refers to MAESTRO 1 study, which has a similar protocol to MAESTRO 3 (unpublished).

Abbreviations: RCT=double-blind randomised controlled trial; RRMS=relapsing-remitting multiple sclerosis; SPMS=secondary progressive multiple sclerosis; PPMS=primary progressive multiple sclerosis; PMID, PubMed Identifier.

Clinical outcomes

The Expanded Disability Status Scale (EDSS)[27], which rates neurological impairment, was scored as per individual study protocol. The EDSS was obtained at least 1-month prior to or after a protocol-defined relapse. We defined disability progression confirmed at 24-week as a worsening of EDSS that was sustained on subsequent visits for at least 24 weeks. EDSS progression was defined as a ≥1.5-point increase from a baseline EDSS of 0, a ≥1-point increase from a baseline score of 0.5 to 5.5, a ≥0.5-point increase from a baseline score greater than 5.5.

Brain MRI Protocol

We included visits where all the following brain MRI sequences were acquired: T1-weighted, T2-weighted, and Fluid Attenuated Inversion Recovery (FLAIR) MRI. The MRI protocol of three cohorts did not include FLAIR scans (see **Supplemental Material** for details).

We used brain 2D or 3D T1-weighted scans to obtain grey and white matter segmentations, FLAIR and T1-weighted scans to segment lesions, and T2-weighted scans, together with T1-weighted scans, to obtain T1/T2 ratio. Details of MRI protocols are explained in previous publications (**Table 1**).

Image processing

We processed MRI scans from MS and healthy volunteers with our pipeline to obtain the following 18 variables according to the Neuromorphometrics atlas (http://www.neuromorphometrics.com):

- Volumes of the bilateral frontal, parietal, temporal, and occipital grey matter, limbic cortex, cerebellar grey matter and white matter, brainstem, deep grey matter and cerebral white matter
- Volume of total T2 lesions
- Regional T1/T2 ratio of normal-appearing white matter in the corpus callosum, frontal, temporal, parietal, and occipital lobes, cingulate bundle and cerebellum.

Details of image analysis and quality control pipelines are explained in detail in the **Supplementary Material**.

Statistical analysis

Outline

In the following steps, we trained, cross-validated, and externally tested our data-driven subtyping and staging machine learning algorithm (SuStaIn) following our previously developed pipeline[10] (**Figure 1**):

1) Adjusting neuroimaging measures for nuisance variables;
2) Calculating Z-score of MRI-derived measures based on the healthy control population;
3) Separating *a priori* MS trials and cohorts into a "train and cross-validation set" and an "external test set";
4) Selecting MRI-derived variables using only the train and cross-validation set;
5) Computational modelling using SuStaIn to select the optimal number of MS subtypes and quantify the uncertainty of the model, by carrying out the leave-one-dataset-out cross-validation;
6) Testing the trained model on the external test set.

Figure 1. Data processing pipeline.

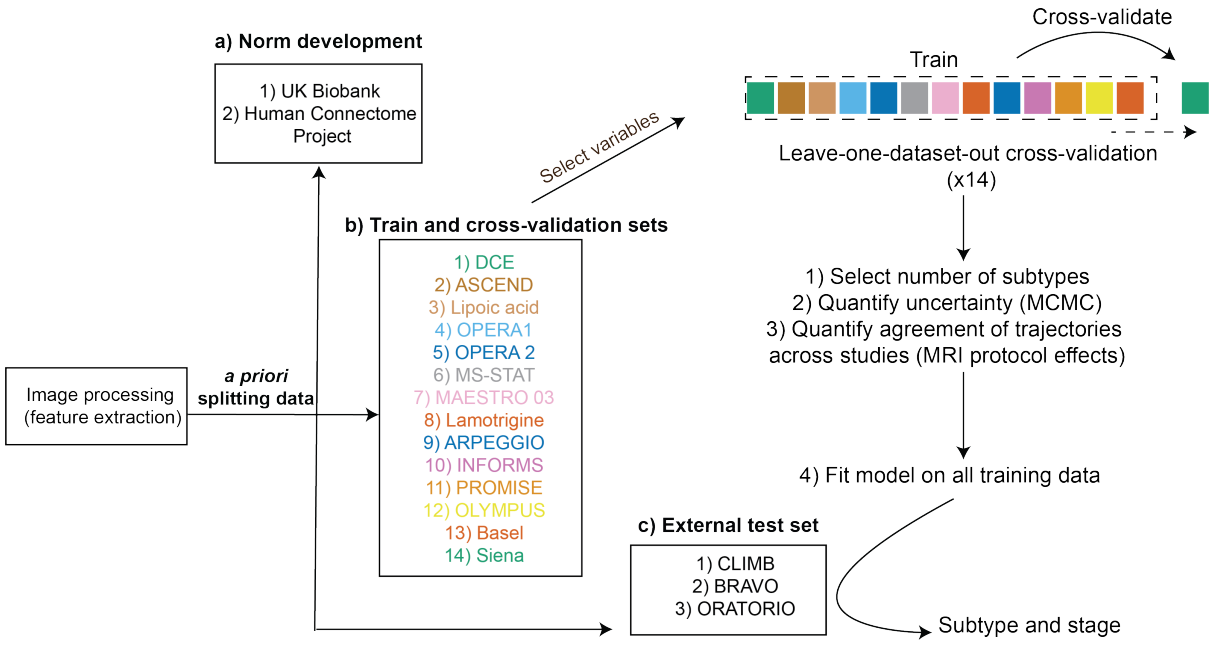

Figure 1 legend: This figure shows that all raw data from different cohorts underwent a unique image processing pipeline to extract neuroimaging variables of lobar grey matter

volume, visible white matter lesion from FLAIR, and T1/T2 ratio. We used the population of healthy volunteers to develop norms, calculate Z-scores, and select MRI variables. *A priori* we split our large patient dataset into two separate datasets: 14 datasets in the "train and cross-validation" sets, and three datasets in the external test set: CLIMB (an observational study), BRAVO (a phase III RRMS trial), and ORATORIO (a phase III PPMS trial). We used the train-cross-validation set to choose the optimal number of subtypes and to quantify the uncertainty. We applied the trained model on the external test set to subtype and stage patients.

Acronyms: MCMC, Markov Chain Monte Carlo; RRMS, relapsing remitting multiple sclerosis; PPMS, primary progressive multiple sclerosis; MRI, magnetic resonance imaging.

We now explain each step below:

1) *Adjusting for nuisance variables: age, sex, total intracranial volume, trial and MRI protocol*

For each of the neuroimaging variables, we used the population of healthy volunteers to fit a Bayesian linear regression model with total intracranial volume, sex, and age as independent variables, and each neuroimaging measure as the outcome variable. We predicted expected values for each subject using this model and subtracted the observed values to obtain residual values of each neuroimaging variable. We refer to the residual values as "adjusted values". We used BAS package version 1.5.3 inside R version 3.6.0[28]. Each trial had a different MRI protocol. We chose not to adjust for MRI protocol inside the abovementioned statistical models because differences in eligibility criteria and differences in MRI protocols would perfectly mirror each other in statistical models and would have regressed out the effects of interest. To ensure that our results were robust to trial-specific nuisance effects, we used a separate method inside the cross-validation (see below for details).

2) Norm development and Z-score calculation

We calculated the Z-score for each MRI variable at each participant's visit by subtracting the adjusted mean value of the healthy volunteers from the adjusted observed value and dividing it by the standard deviation of the healthy volunteers.

3) Creation of the train and cross-validation and test sets

From the 17 datasets available, we *a priori* chose 14 datasets for the train and cross-validation set. These were three phase 3 RRMS trials, three phase 3 PPMS trials, two phase 3 SPMS trial, three phase 2 SPMS trials, one phase 2 PPMS trial, and two observational cohorts (see **Table 1** and **Figure 1** for the complete list).

The remaining three datasets, which were one phase 3 trial in RRMS, one phase 3 trial in PPMS, and one observational cohort in relapase-onset MS, were set aside and used as test set, to ascertain the external generalisability of the newly developed model.

4) Variable selection in the train and cross-validation set

To reduce the dimensionality of the models and computational expenses we selected variables which were more likely to be affected in MS. To do so, we calculated the effect size of the difference between healthy volunteers and patients at their baseline visit in the train and cross-validation dataset and selected those associated with a moderate to large effect size (>0.5 Cohen's D effect size) in pairwise (MS *vs*. HC) comparisons.

5) Computational modelling using SuStaIn

We used Subtype and Stage Inference (SuStaIn)[10], which is an unsupervised machine learning algorithm previously used in Alzheimer's disease and frontotemporal dementia. In short, SuStaIn identifies data-driven subgroups who share patterns of worsening over time in MRI abnormalities by combining unsupervised learning (clustering) and disease progression modelling. SuStaIn extends the "event-based" models, previously used to identify single trajectories of change in MS[29] and other neurodegenerative disorders[30,31]. SuStaIn

simultaneously estimates a set of subtypes, the trajectory (or sequence of progression of MRI abnormalities, described as a Z-score model) of each subtype, and a probabilistic assignment of each subject to a specific subtype and stage within a subtype. Methodological details of the SuStaIn algorithm are explained in detail elsewhere[10].

In this study we used models with three Z-scores (1, 2, and 3) to provide interpretable results on mild, moderate and severe abnormality as patients progress. We entered the variables resulting from the previous step into SuStaIn. Since lower values of volumes and T1/T2 ratios are expected to be associated with increased EDSS scores, we flipped their signs so that higher Z-scores and estimated stages represented disease worsening.

Leave-one-dataset-out cross-validation in the train and cross-validation set

We carried out the leave-one-dataset-out cross-validation, in order to choose the number of MS subtypes, quantify the uncertainty associated with a given subtype "trajectory" (or the evolution of MRI abnormalities), and test the stability and robustness of the model across different trials, field strengths, and MRI protocols. In the leave-one-dataset-out cross-validation, we trained the model on 13 out of the 14 datasets, contained in the train and cross-validation set, and evaluated on the remaining dataset (*held-out* sample). We permuted the training and held-out samples until every dataset was used for evaluation once. The aims of this step were to:

(1) Select the best fitting model, which means the model with the optimal number of subtypes, and sequence of MRI progression in those subtypes. We started by fitting SuStaIn on the training set with only one subtype and then increased the number of subtypes one by one. We calculated the log-likelihood of each fitted model on the held-out sample, and then compared it with the next model with an extra subtype. We chose the best fitting model when log-likelihood of a given model did not show improvement on the held-out sample.

(2) Calculate the uncertainty of the quantification (posterior distribution); we used the Markov Chain Monte Carlo (MCMC) algorithm with 100,000 iterations to sample from the posterior distribution of the most likely sequence found in the previous step. We sampled the posterior distribution across cross-validation folds to quantify the agreement and trial-specific effects.

(3) Quantify the effects of study and MRI protocol on subtype trajectory; we quantified the degree of overlap of posterior distributions of sequences for each subtype across 14 iterations of cross-validation. We used the Bhattacharya coefficient for this purpose between each pair of subtypes from different folds. The Bhattacharya coefficient ranges from 0 (no agreement) to 1 (perfect agreement). We calculated all pairwise Bhattacharya distances across all folds and subtypes pairs for the best model (which was the three-subtype model) and reported the average and standard deviation for each subtype.

(4) Train a unique model on all the train-cross-validation dataset; because each iteration of cross-validation generates a separate "trained" model, we trained a final model, with three subtypes, on all the train-cross-validation dataset to obtain one model for external testing. We also used this model to obtain MCMC samples and visualised the uncertainty of the final model with the posterior distribution of this model.

We performed additional analysis to test the reliability and stability of the SuStaIn subtypes over time (see **Supplemental Material** for details).

6) *Testing SuStaIn on the external, unseen dataset*

We tested our subtyping model on the external test set, which we did not use in the model development. This test set contained the phase III RCT in RRMS (BRAVO)[21], the phase III RCT in PPMS (ORATORIO)[13], and the observational study in relapse-onset MS (CLIMB) [26] (**Table 1**). In the BRAVO RCT and CLIMB study, we performed a survival analysis to calculate the time to 24-week confirmed EDSS progression of each data-driven MS subtype. In the phase 3 RCT in PPMS (ORATORIO), we considered only the placebo arm for the longitudinal

analysis because the time-to-event analysis was modified by the experimental drug[13]. Since the number of 24-week confirmed EDSS progression events per MS subtype in the placebo group was too small (<23 events per subtype on average) to provide statistically reliable results, we defined an event as ≥20% increase in the averaged 9 Hole-Peg Test time between two hands, following Wolinsky et al's recommendation[32].

We used Cox regression models to calculate the hazard ratios of reaching the 24-week confirmed disability progression in the three data-driven MS subtypes. To investigate whether there were differences in disease activity between the data-driven subtypes, we used the annual relapse rate in the BRAVO RCT and CLIMB study (in the ORATORIO trial relapses were extremely rare) and the number of gadolinium-enhancing lesions at baseline in the BRAVO and ORATORIO trials, which was previously reported[13,21] (the CLIMB study which did not include this variable). We used a Poisson model in which the lesion count at baseline in all patients (placebo and treatment arms) was the outcome and the data-driven subtype was the independent variable.

MS clinical classification and EDSS vs. data-driven subtyping and staging

SuStaIn estimates subtype stages along a trajectory or a 'sequence'. Since we had 13 variables with three Z-scores, each subtype had 39 stages: from one (early) to 39 (late). To compare whether the data driven subtypes were associated with disability progression more strongly than clinical MS phenotypes we constructed a mixed-effects model. In this model time to reach 24-week confirmed EDSS progression was the outcome variable and 'trial' was a random-effects variable. Fixed-effects predictors were SuStaIn subtypes and stages at baseline, clinical classifications, age, sex, and EDSS at baseline.

Treatment response in data-driven subtypes

To test whether there is a difference in the treatment response between the SuStaIn subtypes, we looked at the rate (or slope) of EDSS worsening in three progressive MS trials (ORATORIO, ASCEND, and OLYMPUS) pooled together, and in three RRMS trials (DEFINE-

CONFIRM-ENDORSE, OPERA1, and OPERA2) also merged together. We chose them because had at least 100 patients per data-driven subtype, and were either positive trials or had a subgroup that had shown a trend towards a treatment response in previous publications. For the DEFINE-CONFIRM-ENDORSE, we merged the two arms with different doses of the drug, included the placebo arm, and excluded the active comparator arm. We used a linear mixed effects model in which EDSS was the outcome variable with 'group', time, and group-time interaction as the independent variables. 'Group' was a binary variable for each of the data-driven subtypes (placebo or treatment for that subtype). To adjust for repeated measures and correlated residual errors, we added hierarchical random effects to our model, in which visits were nested in the 'subject' variable. We reported the difference between groups in percentage change of EDSS worsening rates across groups, which we refer to as 'treatment response' throughout this manuscript. We used NLME package version 3.1 and Survival package version 2.44 inside R version 3.6.0 for statistical analysis[33,34].

Results

Demographic characteristics of patients and healthy controls

We processed MRI data from 8,545 participants with MS who had a total of 31,451 MRI visits from 17 datasets (**Table 1**). Demographic information was available for 8,241 patients (5,226 women and 3,015 men), whose mean age was 42.31 years (standard deviation 10.5, range: 18 to 69). At baseline visit –defined as the first available clinical assessment– out of 7,923 patients with available clinical information, 4,157 were RRMS, 2,077 PPMS, and 1,689 SPMS.

We also processed MRI scans of 14,928 volunteers (13,823 from the UK Biobank and 1,105 from the Human Connectome Project; 7,965 women and 6,963 men). Their mean age was 52.94 years (standard deviation 9.98, range: 23.5 to 70).

MRI variables selected for the train and cross-validation set

Out of the 18 MRI variables analysed, 13 were associated with a moderate to large effect size when patients were compared with healthy controls, and, therefore were selected and entered into the subtyping algorithm. They were: the volumes of the occipital, parietal, limbic and frontal grey matter, the deep grey matter volume, the total white matter lesion load, and T1/T2 ratio in the corpus callosum, temporal, pariental, cerebellar and frontal normal appearing white matter regions. **Figure 2a** shows the location and effect sizes of the included variables. The remaining 5 variables were excluded.

Three data-driven subtypes best fit the data

The three-subtype model was associated with the highest log-likelihood in the left-out set and therefore considered to be the best model for the data (**Figure 2b**). The two-subtype models showed 11.1% (standard deviation=5.7) increase in log-likelihood compared to one-subtype model and three-subtype models showed 2.1% (standard deviation= 1.9) improvement in log-likelihood than the two-subtype model. The four-subtype model did not show improvement

compared to the three-subtype model (-1.4% change (standard deviation=6.6)). Therefore, we chose three as the optimal number of MS subtypes.

Figure 2. Data-driven variable and subtype selection.

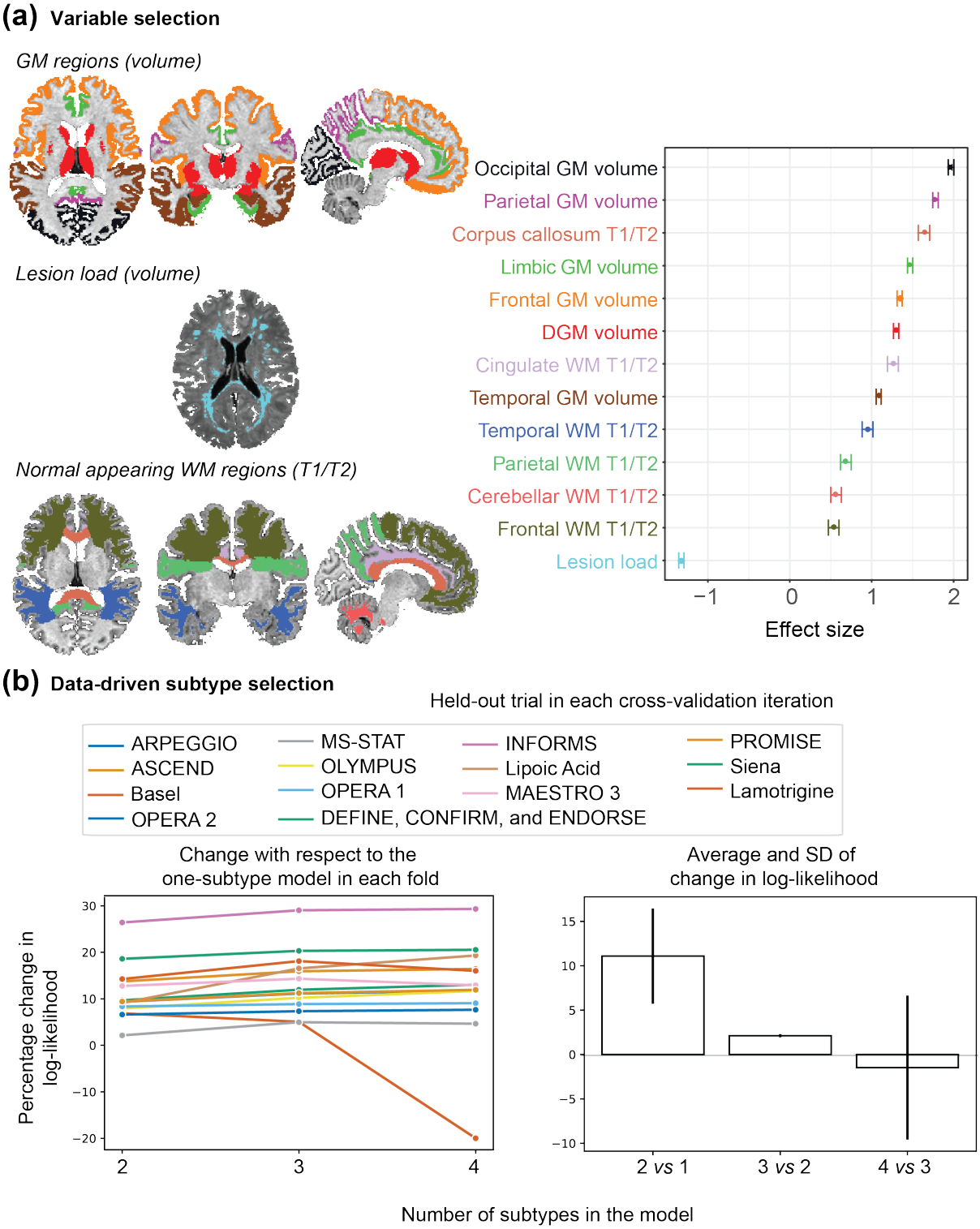

Figure 2 legend: (a) we chose variables whose effect size was medium to large (Cohen's d effect size greater than 0.5) when comparing all patients of the 'train and cross-validation set' with healthy volunteers. We have overlayed selected 13 variables on a randomly chosen

T1-weighted scan of a patient. We used same colour codings to show selected variables on the brain template and the right plot. On the right plot, dots represent point estimates of the effect size and error bars represent the 95% confidence interval of the effect size. (b) We used leave-one-dataset-out in the train and cross-validation set of 14 trials, each time leaving one trial out and fitting SuStaIn algorithm on the remiaing 13 trials. We chose the best number of subtypes according to the log-likelihood of the model calcualted from the *left-out* set each time (x14). The vertical axis shows percentage change in log-likelihood. Larger values favour better models. We found that three subtypes provided the best accuracy (according to log-likelihood) to evaluate subtypes in the lef-out-set.

Acronyms: NAWM, normal-appearing white matter; SD, standard deviation; GM, grey matter; T1/T2, T1-T2 ratio.

When looking at how abnormalities spread across the brain in each subtype, we termed them as follows: (1) The cortex-first subtype (44% of participants, n=3480): starting from the cortical atrophy in the occipital, parietal, and frontal lob GM atrophy before spreading to other areas and preceding lesion accrual, (2) The normal appearing WM (NAWM)-first subtype (30%, n=2369): starting from the reduction in the T1/T2 in the cingulate white matter and the corpus callosum then spreading to the cerebellar white matter abnormality (reduction T1/T1) and late appearance of lesions, and (3) The lesion-first subtype (26%, n=2074) (**Figure 3a** and **Supplementary Figure 1):** starting from severe (sigma> 3) accumulation of lesions, and early deep grey matter atrophy, with a later cortical atorphy. The three subtypes were similar in gender distribution: the cortex-first subtype there were 2182 women (62.6%) and 1299 men, in patients with NAWM-first subtype there were 1552 (65.51%) women and 817 men, and in patients with the lesion-first subtype there were 1311 (63.2%) women and 762 men. Patients in these three data-driven subtypes were similar in age (mean ± standard deviation of age in the cortex-first=41.11±10.13, NAWM-first=42.61±11.04, and lesion first=41.98±10.74). The lesion first subtype had a slightly higher median EDSS at baseline (median=4, 0-9) compared to the other two subtypes (median EDSS for cortex-first 3.5, and NAWM-first=3.5). The lesion first subtype had a longer disease duration (8.03 years) compared to the other two subtypes which was statistically significant (5.4 years for the cortex-first and 5.1 years for the NAWM-first, effect size=0.3 (small), $p<0.01$). The lesion-first subtype had the highest lesion load at baseline ($p<0.001$). The cortex first subtype had the smallest cortex at baseline($p<0.01$). Over time, the lesion-first subtype had the slowest rate of cortical atrophy ($p<0.001$) compared to the cortex-first and NAWM-first subtypes. **Table 2** and **Figure 3b** show these baseline characteristics.

Table 2. Characteristics of clinical and data-driven subtypes at baseline.

Clinical subtypes			
	RRMS	**SPMS**	**PPMS**
Average age (SD)	37.45 (9.30)	49.43 (8.06)	47.72 (8.54)
Percentage female	69%	65%	49%
Median EDSS (interquartile range)	2.5 (1.5 – 3.5)	6 (5 – 6.5)	4.5 (3.5-6)
Average disease duration (SD)	4.19 (5.21)	14.6 (8.81)	3.89 (4.20)
Progression duration	–	5.30 (4.07)	3.89 (4.20)
Data-driven subtypes			
	Cortex-first	**NAWM-first**	**Lesion-first**
Percentage of population	44%	30%	26%
Average age (years, SD)	42.11 (10.13)	42.61 (11.04)	41.98 (10.74)
Percentage female (%)	62%	65%	63%
Median EDSS (interquartile range)	3.5 (0-8)	3.5 (0-7.5)	4 (0-9)
Average disease duration (years, SD)	5.4 (6.5)	5.1 (6.48)	8.03 (8.43)
Average (SD) SuStaIn stage	14.98 (0.1)	14.79 (0.1)	13.81 (0.1)
Lesion volume (± SE, cubic centimetre) *	19.37 ± 0.28	16.03 ± 0.28	31.63 ± 0.30
Cortical volume (± SE, cubic centimetre) *	389 ± 0.22	393 ± 0.23	390 ± 0.23
Percentage annual cortical atrophy (±SE)	2.35% ± 0.05	2.36% ± 0.02	2.10% ± 0.02

Table caption: We used Kruskal-Wallis test for ordinal variables (SuStaIn stages and EDSS), ANOVA for continuous variables (age and disease duration), Chi-square test for gender frequencies, and mixed-effects models for others to estimate *p*-values.

* Predicted marginal means: adjusted for total intracranial volume, age, and expected lesion volume from healthy ageing. Unit is cubic centimetre.

NS=non-significant; SD=standard deviation; SE=Standard error of mean; NAWM=normal-appearing white matter; EDSS=Expanded Disability Status Scale; RRMS=relapsing remitting multiple sclerosis; SPMS=secondary progressive multiple sclerosis, PPMS=primary progressive multiple sclerosis.

In those clinically classified as RRMS, 47.9% were in the cortex-first subtype, 27% in the NAWM-first subtype, and 24.1% were in the lesion-first subtype. In SPMS the proportions in the lesion-first group (40.06%) were similar to the cortex-first subtype (39.1%), with a minority of NAWM-first (20.8%). In PPMS, there was a similar number of cortex-first and NAWM-first subtypes (42.1% and 40.5 %), and a minority (17.4%) in the lesion-first subtype. **Figure 3b** shows distributions between clinical classification and data-driven subtypes.

Figure 3. Sequence of events in data-driven subtypes.

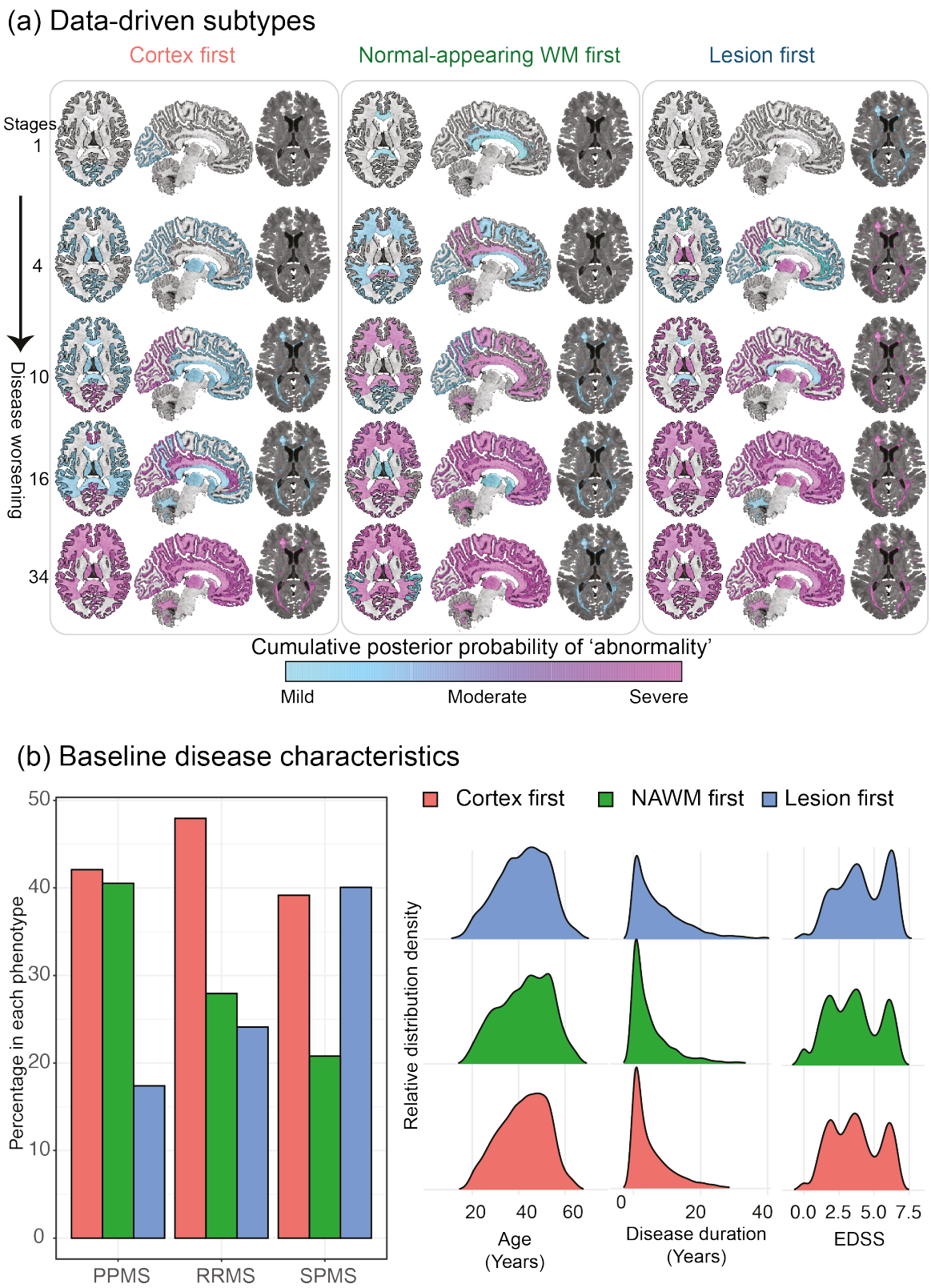

Figure 3 legend: (a) This figure shows how damage spreads across the brain in each of the three data-driven subtypes. The colour shade ranges from blue to pink which represents the

probability of 'abnormality' it can be interpreted as the degree of 'abnormality' (mild, moderate or severe which approximates 1,2 and, 3 Z-score) relative to healthy volunteers. The most common subtype was the cortex-first subtype (left) which has cortical atrophy in the occipital, temporal and frontal cortex at early stages. The normal-appearing white matter (NAWM) damage-first subtype (middle) had early damage in T1/T2 of the corpus callosum and does not show visible white matter lesion until later stages. The lesion-first subtype (right) shows early severe accumulation of lesions, early severe atrophy in the deep gray matter which later spreads to other areas. The numbers on the left side represent SuStaIn stages. The minimum stage is 1 and the maximum stage is 39 (based on 13 variables that can show mild, moderate and severe abnormality; 13x3 = 39). (b) The cortex first subtype was the most common in PPMS and RRMS patients. The lesion-first subtype was the most common in the SPMS patients. The three data-driven subtypes were similar in age, and gender distribution, but the lesion first subtype had statistically significantly higher EDSS and disease duration, as shown in the right side of (b) section of the figure.

MRI protocol and trial-specific effects: agreements of the subtype sequences across trials and cross-validation folds were excellent

The average (±standard deviation) measure of agreements for the three data-driven subtypes (Bhattacharya distance which ranges from not agreement, 0, to perfect agreement 1) of the posterior distribution of the estimated sequence of in each subtype across cross-validation folds were as follows: 0.94 (±0.03) for the cortex-first subtype, 0.94 (±0.02) for the NAWM-first group, and 0.96 (±0.02) for the lesion-first group, suggesting excellent agreement across trials despite different MRI protocols.

Longitudinal staging of patients showed steady worsening over time confirming the reliability of our modelling

When looking at all patients, at baseline the estimated average (±standard error) SuStaIn stage was 14.98 (±0.1) in the cortex-first subtype, 14.79 (±0.1) in the NAWM-first subtype, and

13.81 (±0.1) in lesion-first subtype· There was significant annual increase (p<0.001 for all tests) in stages across patients in the three subtypes, giving confidence to reliability of data-driven staging. Estimated average (±standard error) of annual increase of SuStaIn stage in cortex-first was 0.27 (±0.02), in the NAWM-first was 0.17 (±0.02), and in the lesion-first subtype was 0.20 (±0.02).

External validity and generalizability in unseen datasets: disability progression and disease activity

a) ***Disability progression:*** *the lesion-first subtype had a higher risk of developing 24-week confirmed EDSS progression*

BRAVO

There was a significant difference in reaching 24-week confirmed EDSS progression between the three groups (log-rank test for three-group comparison *p*=0.003). Patients with the lesion-first subtype had 82% (95% confidence interval: 48.6%-115%, p<0.001) higher risk of reaching the 24-week confirmed EDSS progression compared to the cortex-first subtype. There was no difference in reaching the confirmed disability progression between the cortex-first and NAWM-first subtypes. **Figure 4** shows the survival plots for these analyses.

ORATORIO

When looking at all patients (placebo and treatment arms), the lesion-first subtype had the highest EDSS across three subtypes at baseline: model-estimated average difference with the cortex-first subtype was 0.33 (standard error = 0.10, p=0.001) and with the NAWM-first subtype was 0.35 (standard error=0.13, p=0.01). NAWM-first subtype and cortex-first subtype had similar EDSS at baseline (estimated difference=-0.02, standard error =0.12, p=0.8). There were 196 progression events based on ≥20% increase of 9-Hole Peg Test time in the placebo arm. The lesion-first subtype had a significantly shorter time to reach this disability milestone (hazard ratio =2.01, 95% confidence interval=1.33 to 3.33, p=0.007) compared to the cortex-first subtype, but not the NAWM-first subtype (hazard ratio = 2.027, 95% confidence interval=0.96 to 4.26, p=0.06). This means that at any particular time the lesion-first subtype

patients were almost twice (101% more) as likely to reach the disability milestone as the cortex-first subtype.

CLIMB

In the observational CLIMB dataset, there were 212 patients who had sufficient clinical information for this analysis, of whom 111 had 24-week confirmed EDSS progression. Patients with the lesion-first subtype had a significantly shorter time to reach 24-week confirmed EDSS progression compared to the other two subtypes (log-rank test p=0.035). The hazard ratio (95% confidence interval) for reaching 24-week confirmed EDSS progression was 1.75 (1.32-2.18, Cox model regression coefficient p = 0.01) in the lesion-first group *vs.* NAWM-first. Therefore, at any particular time, patients in the lesion-first group had on average 75% (32%-118%) higher risk of reaching 24-week confirmed progression than the NAWM-first subtype. There was no significant difference in the hazard ratio of developing confirmed EDSS progression in the lesion-first subtype *vs*. cortex-first.

b) Disease activity: *data-driven subtyping could predict disease activity*

BRAVO: lesion-first subtype had a higher rate of annual relapse rate and a higher count of gadolinium-enhancing lesions

Estimated average (±95% confidence interval) annual relapse rate for the cortex-first subtype was 0.33 (0.29-0.37), for the NAWM-first was 0.37 (0.32-0.42), and for the lesion-first subtype was 0.43 (0.35-0.51). Thus, the difference between the lesion-first *vs* the cortex-first subtype was statistically significant (p=0.04). When looking at gadolinium-enhancing lesion counts, at baseline in placebo and treatment arms, the lesion-first subtype had an average of 3.44 lesions per subject, which was significantly higher than the cortex-first subtype (Poisson model estimated difference in count with the cortex-first subtype= 146%, standard error = 20.5%, p <0.001). The NAWM-first and the cortex-first subtypes had similar T1-weighted gad-enhancing lesion counts (average of 0.72 and 1.2, p=0.1). **Figure 4C** shows these results.

ORATORIO: there were more gadolinium enhancing lesions in the lesion-first subtype

When looking at gadolinium-enhancing lesions, at baseline in patients (placebo and treatment arms), the lesion-first subtype had an average of 2.25 lesions per subject, which was significantly higher than the cortex-first subtype (Poisson model estimated difference in count with the cortex-first subtype= 429%, standard error = 109%, p <0.01). The NAWM-first subtype and the cortex-first subtype had similar T1-weighted gadolinium-enhancing lesion counts (average of 0.42 and 0.21, p=0.3). Relapses were too rare to provide statistically reliable results. **Figure 4** shows these results.

Data-driven subtyping and staging could better predict 24-week confirmed EDSS progression than clinical phenotypes or baseline EDSS

There were significant associations (standardised β) between data-driven subtypes (overall subtype effect, β = 0.04, standard error = 0.01, p = 0.02) or data-driven baseline stages (β = -0.06, standard error=0.02, p<0.001) with the time to reach 24-week confirmed EDSS progression. However, there were no significant associations of clinical classification of MS (overall effect across RR, SP and PP MS β =0.18, standard error = 0.15, p=0.22) or baseline EDSS (β =0.02, standard error = 0.03, p=0.26) with time to reach 24-week EDSS progression in the same model, suggesting that data-driven subtyping and staging has a better predictive value for clinical progression, than EDSS at baseline or clinical phenotypes.

Treatment response

Progressive MS

Patients in the lesion-first subtype had the largest slowing in EDSS worsening in pooled dataset of anti-inflammatory drug trials in progressive MS

In our pooled analysis of ORATORIO, ASCEND, and OLYMPUS the percentage change in EDSS worsening of each data-driven subtype with respect to the same subtype on treatment was as follows (each subtype on treatment *vs.* same subtype on placebo; model estimated percentage average difference ± standard error; negative values represent slowing of EDSS worsening):

- Lesion first: -66% (±25.6%), p=0.009
- NAWM-first: -22% (±25.06%), p=0.06
- Cortex-first: 27.6% (±20.17%), p=0.7

This suggests that the lesion-first subtype had the greatest treatment response**. Figure 4d** shows these results.

RRMS

In our pooled analysis of OPERA1, 2, and DEFINE/CONFIRM/ENDORSE, the percentage change of EDSS worsening for each subtype on treatment *vs* placebo (or active comparator) was as follows:

- Lesion-first: -89% (± 44%), p=0.04
- NAWM-first: -75% (±198%), p=0.74
- Cortex-first: 63% (±164%), p=0.70

This suggests that the lesion-first subtype had the greatest treatment response, irrespective of clinical phenotypes of MS.

Longitudinal subtyping: the majority of patients did not change subtype over time providing confidence in our data-driven subtyping

When we looked at 5,973 subjects who were assigned to one of the three data-driven subtypes with probability of more than 95% at study entry (baseline), during the follow up only 8% changed subtype. When we increased the probability threshold to 99% (4,525 subjects at baseline), only 6% of subjects changed subtype over time (94% stayed in the same subtype). When including all subjects without a threshold of probability in subtyping, 25% of patients changed subtype over time.

Staging and stratification

Patients with higher SuStaIn stage at baseline had a faster EDSS progression

Patients who had the highest tertile of SuStaIn stage at baseline (17 to 39) had the shortest time to 24-week confirmed EDSS progression (independent of subtype, log-rank p<0.0001). Average (95% confidence interval) risk of developing 24-week confirmed progression at any

particular time was 37% (22-53%) higher ($p<0.0001$) in this group with respect to the first tertile (stages 1 to 9) and 30% (17 to 46%) higher ($p<0.0001$) in this group with respect to the second tertile group (stages 10 to 17). Our results did not differ when we repeated this analysis inside each data-driven subtype. **Figure 4d** shows these results.

Figure 4. External generalisability.

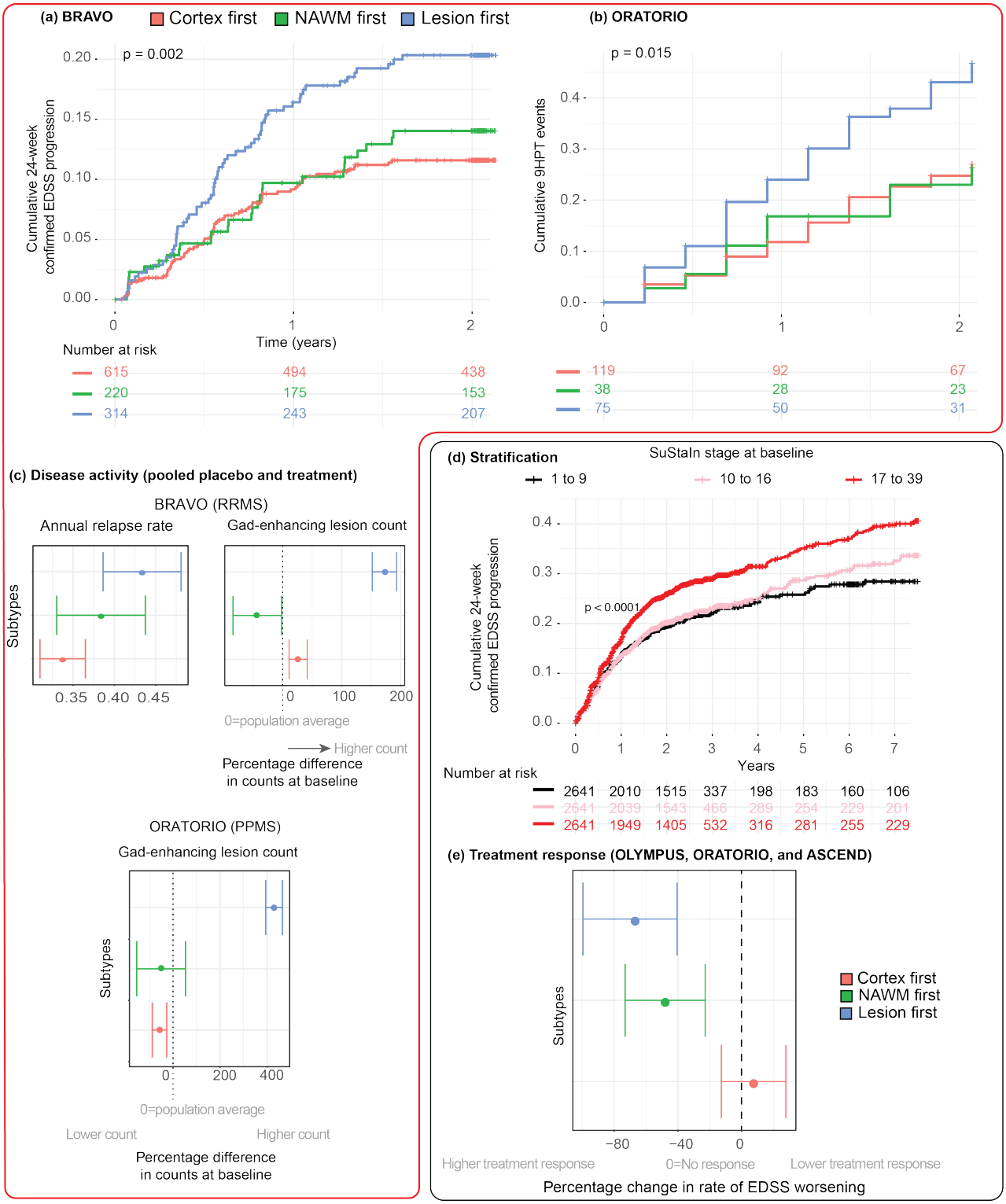

Figure 4 legend: this figure shows our analyses in the external test sets (red border) and pooled datasets (black border). (a) Patients in the lesion-first subtypes had a faster EDSS progression in BRAVO; (b) patients with lesion-first subtype had faster 9-Hole Peg Test progression (see the main text for definition) compared to the other two subtypes in the placebo arm of the ORATORIO trial (we did not use treatment

arm to look at longitudinal changes here because it was a positive trial). The vertical axis shows the 24-week confirmed EDSS progression; (c) top diagram: the lesion-first subtype had more annual relapse rates when looking at all patients in the BRAVO trial than the NAWM-first or cortex-first subtypes; bottom diagram: the lesion-first subtype had more gadolinium enhancing lesions in all patients of ORATORIO trial at baseline; (d) this figure shows that patients who were in the higher tertile of SuStaIn stages had a shorter time to progression: the higher the stage at baseline, the shorter the time to reach 24-week confirmed EDSS progression. When we repeated this analysis inside each subtype we found similar results (not shown); (e) Shows the change in EDSS worsening in data-driven subtypes in the pooled treatment arms of the ORATORIO, ASCEND and OLYMPUS trials compared to the corresponding subtypes in the pooled placebo arms (e.g., lesion-first subtype on treatment *vs.* lesion first subtype on placebo and so forth). Patients in the lesion-first subtype had the largest reduction in the rate of EDSS worsening, which was numerically lower than the normal appearing white matter damage first subtype, and statistically significantly lower than the cortex first subtype.

Acronyms: 9HPT, 9-Hole Peg test; NAMW, normal-appearing white matter; EDSS, Expanded Disability Status Scale; RRMS, relapsing-remitting multiple sclerosis; PPMS, primary progressive multiple sclerosis.

Discussion

To our knowledge this is the first data-driven subtyping of MS to demonstrate significantly better predictive value for treatment response and disability progression than clinical phenotypes. Our subtypes showed high reliability across 14 datasets and were robust to nuisance effects of different MRI protocols. We tested our model in three external datasets and with variables that we did not use in training our model (relapses, gadolinium-enhanicing lesions, EDSS worsening, and Nine-hole Peg Test performance). We found that subtyping of patients could predict disability progression, treatment response, and disease activity (relapse rate and gadolinium-enhancing lesions). We stratified patients according to the SuStaIn stage, we found that higher stages at baseline could predict faster disability progression. SuStaIn provided consistent results across trials and when applied longitudinally suggesting excellent reliability; when we subtyped patients over time, 94% of patients who were assigned with high certainty to a subtype stayed in the same subtype. Our subtyping model has the potential to be used prospectively to enrich clinical trials.

A striking finding was the absence of one-to-one correspondence between data-driven subtypes and clinical MS phenotypes. This confirms previous observations that clinical MS phenotypes do not represent distinct pathological entities[3,4,35]. Even within the same MS clinical phenotype, SuStaIn could identify patients who were more likely to clinically progress during follow-up (e.g., RRMS patients in the BRAVO trial or PPMS patients in the ORATORIO), and progressive MS patients who were most likely to respond to treatment (OLYMPUS, ORATORIO, and ASCEND), and RRMS patients who were most likely to respond to treatments (OPERA1, OPERA2, and DEFINE-CONFIRM-ENDORSE). Since our data-driven subtyping only requires data from a single visit, it has the potential to prospectively enrich trials with those most likely to respond to treatment.

The lesion-first subtype was the least common (26% overall), a slightly higher proportion was classified as NAWM-first (30%), and a more substantial proportion as cortex-first (44%). Counterintuitively, in RRMS, a cortex-first subtype was more common than a lesion-first subtype (47.9% and 24.1% respectively) but in SPMS the proportion classified as having a lesion-first subtype was substantially higher (40.06% and 39.1% respectively). However, given that not all those with RRMS will develop SPMS, a plausible explanation is that the lesion-first subtype is more likely to be associated with disability progression. Therefore, our RRMS group represents a combination of those less likely (cortex-first) and more likely (lesion-first) to eventually develop SPMS. The lesion first subtype had a faster EDSS progression, which further supports these findings.

Our findings provide insights into the sequence and focus of pathology, albeit with the caveat that MRI measures are not pathologically specific. The most common MRI phenotype was the cortex-first with faster cortical atrophy rates but lower lesion volumes. Cortical atrophy in this subtype started posteriorly and spread forward, increasingly involving deep grey matter, with abnormalities in NAWM occurring late and with relatively modest WM lesion accumulation. This suggests a pathological process that is predominantly in the cortex rather than WM-based, dominated by posterior and then more extensive neurodegeneration (as reflected by atrophy)[36,37]. In contrast, the lesion-first subtype starts with marked accumulation of lesions, severe early atrophy in the deep grey matter structures, and then cortical atrophy, with NAWM abnormalities as a late feature. This suggests a process more closely linked with WM lesion accrual, and these findings are consistent with neurodegeneration in deep grey matter secondary to lesion accumulation [36,37]. This may extend previous findings that cortical abnormalities may also be the result of WM lesion effects on tracts[38,39]. The sequence of the NAWM-first subtype suggests a more diffuse process, with limited WM lesion formation even by the late stages, and subsequent GM abnormalities detectable anteriorly and then posteriorly. It is worth noting that patients could belong to multiple MRI-driven subtypes simultaneously, although most (75%) showed a clear dominance of one subtype (defined by probabilistic assignment to each subtype in SuStaIn models). In part this may be due to the

precision with which SuStaIn could subtype people based on the available data, but it also raises the possibility that for some people with MS, multiple pathological mechanisms may contribute to disease progression.

Our study was not without limitations. Different MRI protocols from different trials could limit our conclusions. However, to mitigate the effects of MRI protocol we repeated our analysis using leave-one-dataset out cross validation and found excellent agreement between trajectories (sequence of progression of abnormality) between the three subtypes. Moreover, when we applied our model in external datasets, which had different protocols, we found consistent results which proves that our model is robust to different protocol effects. Secondly, spinal cord is affected from early stages in MS and its atrophy is associated with disability[40]. However, spinal cord data is not routinely acquired in MS trials and was not available in our study. Future studies with spinal cord data should investigate whether spinal cord measures can contribute to SuStaIn subtyping and staging, and whether or not they do so independently or concurrently with brain atrophy.

MRI data-driven subtyping can predict the future clinical course of MS based on a single set of scans. Data-driven subtyping has the potential to be used prospectively to enrich clinical trials with patients who are more likely to respond to treatments.

Acknowledgments

This study was supported (in part) by (an) award(s) from the International Progressive MS Alliance, award reference number PA-1603-08175. We are grateful to all the investigators who have contributed trial data to this study (see the list of investigators in the appendix). We are grateful to Professor Geraint Rees for his comments. We thank Rozie Arnaoutellis, Istvan Morocoz, and Caramanos Zografos for coordinating and organising this study. We thank Jonathan Steel for IT support during this work. This research in part has been conducted using the UK Biobank Resource under Application Number 47233. Data have also been provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. O.C., D.C, A.J.T., F.B., and DCA have received funding from the National Institute for Health Research University College London Hospitals Biomedical Research Centre for this work. D.C.A. has received funding for this work from Engineering and Physical Sciences Research Council Grants M020533, M006093, and J020990, as well as the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements 634541 and 666992.

Disclousres

The authors have no competing interesting with respect to this research. The full disclosure statement is as follows:

AE has received speaker's honoraria from Biogen and At The Limits educational programme. He has received travel support from the National Multiple Sclerosis Society. DC has received honoraria (paid to his employer) from Excemed (2017) for faculty-led education work; had meeting expenses funded by the IMSCOGS (2019), EAN (2018), ECTRIMS (2018) and Société des Neurosciences (2017). He has received research funding from the International Progressive MS Alliance, the MS Society, and the National Institute for Health Research (NIHR) University College London Hospitals (UCLH) Biomedical Research Centre. He is a member of the MS Society's Biomedical Grant Review Panel and a trustee of the MS Trust. OC has received research grants from the MS Society of Great Britain & Northern Ireland, National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre, EUH2020, Spinal Cord Research Foundation, and Rosetrees Trust. She serves as a consultant for Novartis, Teva, and Roche and has received an honorarium from the American Academy of Neurology as Associate Editor of Neurology and serves on the Editorial Board of Multiple Sclerosis Journal. CRGG has received research grants form Sanofi and the National Multiple Sclerosis Society. F.B has received compensation for consulting services and/or speaking activities from Bayer Schering Pharma, Biogen Idec, Merck Serono, Novartis, Genzyme, Synthon BV, Roche, Teva, Jansen research and IXICO and is supported by the NIHR Biomedical Research Centre at UCLH. AJT has received honoraria/support for travel for consultancy from Eisai, Hoffman La Roche, Almirall, and Excemed, and support for travel for consultancy as chair of the International Progressive MS Alliance Scientific Steering Committee, and from the National MS Society (USA) as a member of the Research Programs Advisory

Committee. He receives an honorarium from SAGE Publishers as Editor-in-Chief of Multiple Sclerosis Journal and a free subscription from Elsevier as a board member for the Lancet Neurology. DLA has received research grant funding and/ or personal compensation for consulting from Acorda, Adelphi, Alkermes, Biogen, Celgene, Frequency Therapeutics, Genentech, Genzyme, Hoffman-La Roche, Immuene Tolerance Network, Immunotec, MedDay, EMD Serono, Novartis, Pfizer, Receptos, Roche, Sanofi-Aventis, Canadian Institutes of Health Research, MS Society of Canada, and International Progressive MS Alliance; and holds an equity interest in NeuroRx Research. SN has served on an advisory board for Genentech, has received travel funding from MedDay, research funding from Immunotec and is a part-time employee of NeuroRx Research. DCA, FP, PW, and AY have nothing to disclose.

References

1 Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. *Neurology* 1996; **46**: 907–11.

2 Lublin FD, Reingold SC, Cohen JA, *et al.* Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology* 2014; **83**: 278–86.

3 Stys PK, Zamponi GW, van Minnen J, Geurts JJG. Will the real multiple sclerosis please stand up? *Nat Rev Neurosci* 2012; **13**: 507–14.

4 Thompson AJ, Baranzini SE, Geurts J, Hemmer B, Ciccarelli O. Multiple sclerosis. *The Lancet* 2018; **391**: 1622–36.

5 Fox RJ, Coffey CS, Conwit R, *et al.* Phase 2 Trial of Ibudilast in Progressive Multiple Sclerosis. *N Engl J Med* 2018; **379**: 846–55.

6 Cadavid D, Mellion M, Hupperts R, *et al.* Safety and efficacy of opicinumab in patients with relapsing multiple sclerosis (SYNERGY): a randomised, placebo-controlled, phase 2 trial. *Lancet Neurol* 2019; **18**: 845–56.

7 Hawker K, O'Connor P, Freedman MS, *et al.* Rituximab in patients with primary progressive multiple sclerosis: results of a randomized double-blind placebo-controlled multicenter trial. *Ann Neurol* 2009; **66**: 460–71.

8 Filippi M, Brück W, Chard D, *et al.* Association between pathological and MRI findings in multiple sclerosis. *Lancet Neurol* 2019; **18**: 198–210.

9 Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. *N Engl J Med* 2018; **379**: 1452–62.

10 Young AL, Marinescu RV, Oxtoby NP, *et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* 2018; **9**: 4273.

11 Wolinsky JS, Narayana PA, O'Connor P, *et al.* Glatiramer acetate in primary progressive multiple sclerosis: results of a multinational, multicenter, double-blind, placebo-controlled trial. *Ann Neurol* 2007; **61**: 14–24.

12 Lublin F, Miller DH, Freedman MS, *et al.* Oral fingolimod in primary progressive multiple sclerosis (INFORMS): a phase 3, randomised, double-blind, placebo-controlled trial. *Lancet Lond Engl* 2016; **387**: 1075–84.

13 Montalban X, Hauser SL, Kappos L, *et al.* Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *N Engl J Med* 2017; **376**: 209–20.

14 Kapoor R, Ho P-R, Campbell N, *et al.* Effect of natalizumab on disease progression in secondary progressive multiple sclerosis (ASCEND): a phase 3, randomised, double-blind, placebo-controlled trial with an open-label extension. *Lancet Neurol* 2018; **17**: 405–15.

15 Chataway J, Schuerer N, Alsanousi A, *et al.* Effect of high-dose simvastatin on brain atrophy and disability in secondary progressive multiple sclerosis (MS-STAT): a randomised, placebo-controlled, phase 2 trial. *Lancet Lond Engl* 2014; **383**: 2213–21.

16 Kapoor R, Furby J, Hayton T, *et al.* Lamotrigine for neuroprotection in secondary progressive multiple sclerosis: a randomised, double-blind, placebo-controlled, parallel-group trial. *Lancet Neurol* 2010; **9**: 681–8.

17 Spain R, Powers K, Murchison C, *et al.* Lipoic acid in secondary progressive MS: A randomized controlled pilot trial. *Neurol Neuroimmunol Neuroinflammation* 2017; **4**: e374.

18 Freedman MS, Bar-Or A, Oger J, *et al.* A phase III study evaluating the efficacy and safety of MBP8298 in secondary progressive MS. *Neurology* 2011; **77**: 1551–60.

19 Gold R, Kappos L, Arnold DL, *et al.* Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *N Engl J Med* 2012; **367**: 1098–107.

20 Hauser SL, Bar-Or A, Comi G, *et al.* Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med* 2017; **376**: 221–34.

21 Vollmer TL, Sorensen PS, Selmaj K, *et al.* A randomized placebo-controlled phase III trial of oral laquinimod for multiple sclerosis. *J Neurol* 2014; **261**: 773–83.

22 Miller KL, Alfaro-Almagro F, Bangerter NK, *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* 2016; **19**: 1523–36.

23 Van Essen DC, Smith SM, Barch DM, *et al.* The WU-Minn Human Connectome Project: an overview. *NeuroImage* 2013; **80**: 62–79.

24 De Stefano N, Stromillo ML, Giorgio A, *et al.* Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2015; : jnnp-2014-309903.

25 Bendfeldt K, Hofstetter L, Kuster P, *et al.* Longitudinal gray matter changes in multiple sclerosis--differential scanner and overall disease-related effects. *Hum Brain Mapp* 2012; **33**: 1225–45.

26 Rotstein DL, Healy BC, Malik MT, Chitnis T, Weiner HL. Evaluation of no evidence of disease activity in a 7-year longitudinal multiple sclerosis cohort. *JAMA Neurol* 2015; **72**: 152–8.

27 Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983; **33**: 1444–52.

28 Merlise Clyde, Vandenman, Raj A. Merliseclyde/Bas: Bas Version 1.5.3. Zenodo, 2018 DOI:10.5281/zenodo.1475297.

29 Eshaghi A, Marinescu RV, Young AL, *et al.* Progression of regional grey matter atrophy in multiple sclerosis. *Brain J Neurol* 2018; published online May 8. DOI:10.1093/brain/awy088.

30 Young AL, Oxtoby NP, Daga P, *et al.* A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain J Neurol* 2014; **137**: 2564–77.

31 Oxtoby NP, Young AL, Cash DM, *et al.* Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain J Neurol* 2018; **141**: 1529–44.

32 Wolinsky JS, Montalban X, Hauser SL, *et al.* Evaluation of no evidence of progression or active disease (NEPAD) in patients with primary progressive multiple sclerosis in the ORATORIO trial. *Ann Neurol* 2018; **84**: 527–36.

33 R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2014 http://www.R-project.org/.

34 Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. 2017. https://cran.r-project.org/web/packages/nlme/citation.html (accessed Oct 18, 2017).

35 Coetzee T, Thompson AJ. Unified understanding of MS disease is required for drug development. *Nat Rev Neurol* 2018; published online Jan 12. DOI:10.1038/nrneurol.2017.184.

36 Calabrese M, Magliozzi R, Ciccarelli O, Geurts JJG, Reynolds R, Martin R. Exploring the origins of grey matter damage in multiple sclerosis. *Nat Rev Neurosci* 2015; **16**: 147–58.

37 Geurts JJG, Calabrese M, Fisher E, Rudick RA. Measurement and clinical effect of grey matter pathology in multiple sclerosis. *Lancet Neurol* 2012; **11**: 1082–92.

38 Bodini B, Chard D, Altmann DR, *et al.* White and gray matter damage in primary progressive MS: The chicken or the egg? *Neurology* 2016; **86**: 170–6.

39 Kolasinski J, Stagg CJ, Chance SA, *et al.* A combined post-mortem magnetic resonance imaging and quantitative histological study of multiple sclerosis pathology. *Brain J Neurol* 2012; **135**: 2938–51.

40 Reich DS, Lucchinetti CF, Calabresi PA. Multiple Sclerosis. *N Engl J Med* 2018; **378**: 169–80.

41 De Stefano N, Stromillo ML, Giorgio A, *et al.* Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2016; **87**: 93–9.

42 Kapoor R, Ho P-R, Campbell N, *et al.* Effect of natalizumab on disease progression in secondary progressive multiple sclerosis (ASCEND): a phase 3, randomised, double-blind, placebo-controlled trial with an open-label extension. *Lancet Neurol* 2018; **17**: 405–15.

43 Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 2007; **5**: 11–34.

44 Gorgolewski K, Burns CD, Madison C, *et al.* Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinformatics* 2011; **5**: 13.

45 Klein A, Tourville J. 101 labeled brain images and a consistent human cortical labeling protocol. *Front Neurosci* 2012; **6**. DOI:10.3389/fnins.2012.00171.

46 Eshaghi A, Prados F, Brownlee WJ, *et al.* Deep gray matter volume loss drives disability worsening in multiple sclerosis. *Ann Neurol* 2018; **83**: 210–22.

47 Eshaghi A, Kievit RA, Prados F, *et al.* Applying causal models to explore the mechanism of action of simvastatin in progressive multiple sclerosis. *Proc Natl Acad Sci U S A* 2019; **116**: 11020–7.

48 Tustison NJ, Avants BB, Cook PA, *et al.* N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 2010; **29**: 1310–20.

49 Schmidt P, Gaser C, Arsic M, *et al.* An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage* 2012; **59**: 3774–83.

50 Kamnitsas K, Ledig C, Newcombe VFJ, *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 2017; **36**: 61–78.

51 Prados F, Cardoso MJ, Kanber B, *et al.* A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. *NeuroImage* 2016; **139**: 376–84.

52 Cardoso MJ, Modat M, Wolz R, *et al.* Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion. *IEEE Trans Med Imaging* 2015; **34**: 1976–88.

53 Righart R, Biberacher V, Jonkman LE, *et al.* Cortical pathology in multiple sclerosis detected by the T1/T2-weighted ratio from routine magnetic resonance imaging. *Ann Neurol* 2017; published online Aug 18. DOI:10.1002/ana.25020.

54 Glasser MF, Coalson TS, Robinson EC, *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* 2016; **536**: 171–8.

55 Glasser MF, Van Essen DC. Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *J Neurosci Off J Soc Neurosci* 2011; **31**: 11597–616.

Supplementary Figure 1. Positional variance diagram of three data-driven subtypes of multiple sclerosis.

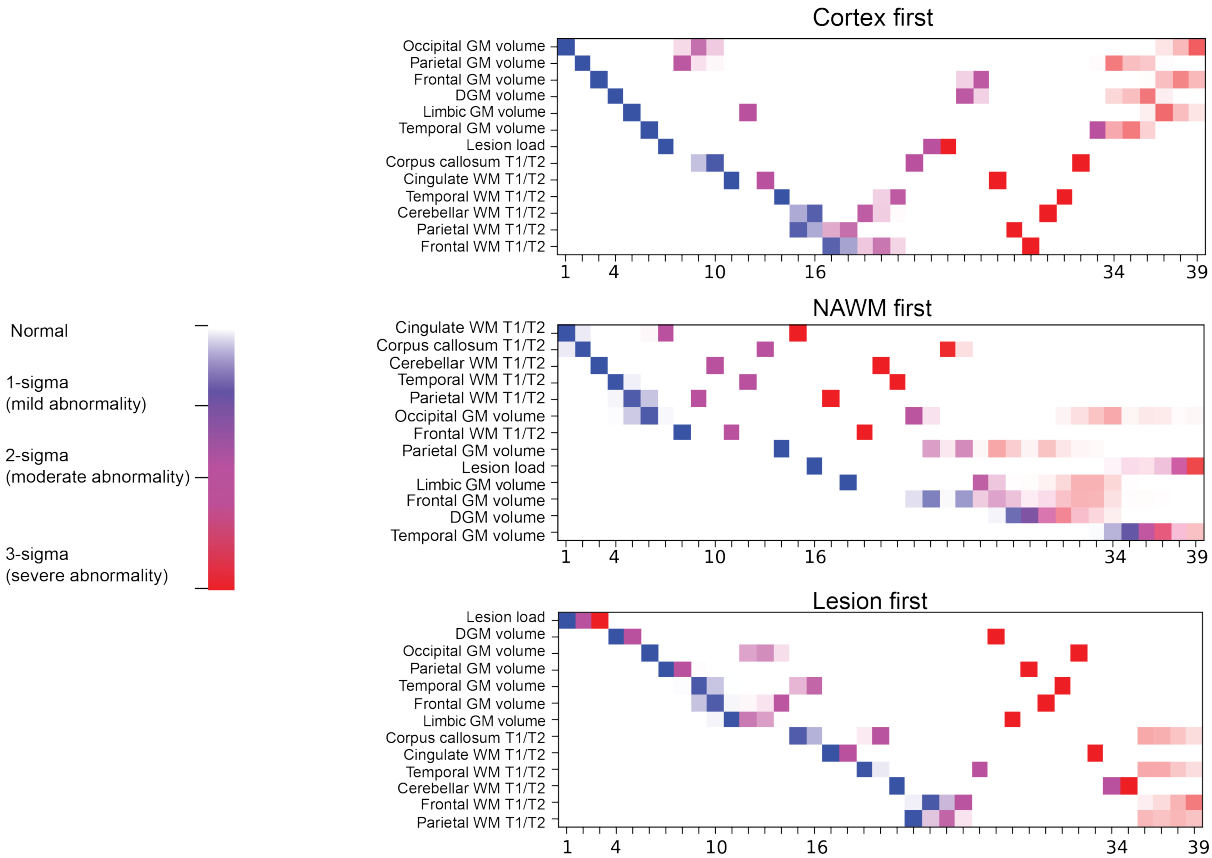

Figure legend: positional variance diagram for three data-driven imaging subtypes of multiple sclerosis. The three different colours represent the degree of abnormality based on Z-score (sigma or standard deviation) models: mild=blue, moderate=violet, and severe=red. The colour shades represent the uncertainty associated with each event position in the posterior distribution of 100,000 Markov Chain Monte Carlo samples.

Acronyms: DGM, deep grey matter; T1/T2, T1-T2 ratio; WM, white matter; GM, grey matter; DGM, deep grey matter; NAWM, normal-appearing white matter.

Supplemental material

Brain MRI Protocol

MRI protocols differed between trials but inside each trial a unique MRI protocol has been used (except for the CLIMB study). We included brain 2D or 3D T1-weighted, fluid attenuated inversion recovery (FLAIR), and T2-weighted MRI scans. **Supplementary Table 1** shows the list of included trials with corresponding publications that reported details of MRI protocol.

In the CLIMB study, in which the MRI protocol had changed over time from 2D T1-weighted MRI to 3D, we only included more recent 3D T1 weighted MRI data.

Supplmental Methods

Image analysis

Brain MRI data handling

We checked and labelled the sequence of MRI scans by visually inspecting nine slices of each MRI scan (three axial, three sagittal, and three coronal slices) with equal slice intervals from the coordinates of the "centre of gravity" of each scan. We organised and uploaded MRI data to an XNAT server (version 1.7.4)[43]. We implemented our image analysis pipeline inside XNAT with Nipype version 1.1.4 to enable large-scale high-throughput computing[44].

Regional brain volume calculation

We aimed to analyse scans to extract volumes of the grey matter regions according to an established brain atlas developed by Klein and Tourville (Neuromorphometrics, http://www.neuromorphometrics.com, see above for the list)[45]. We applied an identical cross-sectional pipeline (treating each visit independently) to all the visits of patients and healthy controls in which T1-weighted, FLAIR and T2-weighted MRI were available. We chose a cross-sectional, rather than a longitudinal image processing pipeline, to ensure that our subtyping models can be used prospectively in the real-world datasets in which (future) follow up data are not yet acquired. We adapted our established MRI analysis pipeline, which we had previously validated in clinical trials and observational cohorts as explained elsewhere in detail[46,47]. Briefly, it included intensity inhomogeneity correction of the T1-weighted MRI with ITK version 5.0 N4-bias field correction algorithm[48], automatic segmentation of hyperintense

lesions of the FLAIR sequence using the consensus (intersection) mask of two different methods (the regression based method in Lesion Segmentation Toolbox version 2.0.15[49] and a deep convolutional neural network based method in DeepMedic version 0.7.1[50]), rigid registration of FLAIR to T1-weighted MRI with co-registration of the FLAIR lesion masks to T1-weighted MRI using ANTs version 2.1.0, and lesion filling with NiftySeg version 1.0[51]. We segmented and parcellated the brain into Neuromorphometrics atlas regions on lesion-filled T1-weighted scans using the Geodesic Information Flows (GIF) software version 3.0[52]. We used a modified version of this pipeline for the Siena cohort, ARPEGGIO and lamotrigine trials which did not have FLAIR but whose investigators had provided manually delineated lesion masks.

T1/T2 ratio calculation of the normal-appearing white matter regions

Lesion masks or brain volumes do not provide any quantitative information on microstructural changes in the white matter. We therefore chose T1/T2 ratio as a measure of extra-lesional white matter changes, because T1 and T2-weighted MRI are widely available in clinical trials and clinical practice (as opposed to more advanced MRI sequences such as diffusion imaging or magnetisation transfer ratio). T1/T2 ratio is an extensively used measures of microstructural changes[53,54]. We adapted available pipelines from the Human Connectome Project to calculate T1/T2 ratio maps for all trials[55]. We corrected for intensity inhomogeneity in T1 and T2-weighted MRI scans with N4 bias field correction algorithm. Next, we rigid-registered T1 and T2-weighted scans in a symmetric space, such that both modalities equally underwent only one interpolation to minimise interpolation artefacts. We calculated the T1/T2 ratio and normalised its value against the average T1/T2 ratio in the ventricles with the co-registered ventricular masks obtained from the GIF segmentation (explained above). We extracted T1/T2 ratio from bilateral normal-appearing white matter regions (see above for list of regions) after we removed co-registered lesions

segmented in FLAIR from the white matter regions, which we refer to as normal appearing T1/T2 ratio throughout this manuscript. Since the T1/T2 ratio in the grey matter regions were highly correlated with grey matter volumetric results, we did not include any T1/T2 ratio in the grey matter in our models.

Quality control

We developed a pipeline to check the quality of results of our pipeline by automatically generating 18 images from segmentation results, lesion segmentations, and registration results which we manually reviewed. We re-ran image analysis pipeline where we identified mis-registrations or faulty segmentations. We did not exclude any visit in clinical trials to perform an intention-to-treat analysis in individuals who met the minimal MRI criteria (which was availability of T1-, T2-weighted, and FLAIR).

Nipype codes for our image analysis and quality assessment pipelines can be accessed under an open licence at https://github.com/armaneshaghi/nipype_codes_ipmsa .

Supplemental Statistical Analysis

Reliability and stability of SuStaIn models: longitudinal subtyping

In addition to subtyping patients at baseline, we trained our model on the baseline subjects and predicted the probability of subtype membership for the available patient visits over time (31,451 visits). We reported the number of subjects who preserved the subtype membership. To calculate the annual rate of change in SuStaIn stages for each data-driven subtype, we fit a mixed-effects model in which the SuStaIn stage was an outcome variable and time was the independent variable (fixed effects). In these models to adjust for hierarchical repeated measures, we defined nested random effects in which 'time' variable was nested in the 'subject' variable. To calculate longitudinal cortical atrophy in each subtype we used a similar

mixed effects model and log-transformed the cortical volumes to obtain the annual percentage volume change.

International Progressive MS Alliance (PMSA) Investigators

Name	Institution
Douglas L Arnold	McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
Sridar Narayanan	McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
Frederik Barkhof	Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, WC1B5EH, UK
Olga Ciccarelli	Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, WC1B5EH, UK
Declan Chard	Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain

	Sciences, University College London, WC1B5EH, UK
Louis Collins	McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
Tal Arbel	McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
Charles R.G Guttman	Center for Neurological Imaging, Brigham and Women's Hospital, Harvard Medical School, Massachusetts, USA
Jerry S Wolinsky	McGovern Medical School, The University of Texas Health Science Center at Houston (UTHealth), Houston, Texas, USA
Garry R Cutter	University of Alabama at Birmingham School of Public Health, USA
Nicola De Stefano	University of Siena, Italy
Maria Pia Sormani	University of Genoa, Italy
Ludwig Kappos	University Hospital Basel, Switzerland
Jack H Simon	Oregon Health and Sciences University, Portland Veterans Affairs Medical Center, Oregon, USA
Jeremy Chataway	Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, WC1B5EH, UK
Raj Kapoor	Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, UCL Queen Square Institute of Neurology, Faculty of Brain

	Sciences, University College London, WC1B5EH, UK
Howard L. Weiner (CLIMB investigator)	Brigham and Women's Hospital, Ann Romney Center for Neurologic Diseases, Department of Neurology, Boston, MA, 02115
Tanuja Chitnis (CLIMB investigator)	Brigham and Women's Hospital, Ann Romney Center for Neurologic Diseases, Department of Neurology, Boston, MA, 02115
Rohit Bakshi (CLIMB investigator)	Brigham and Women's Hospital, Ann Romney Center for Neurologic Diseases, Department of Neurology, Boston, MA, 02115