frontiers | Frontiers in Radiology

TYPE Original Research
PUBLISHED 11 November 2022
DOI 10.3389/fradi.2022.1026442

OPEN ACCESS

EDITED BY
Sotirios Bisdas,
University College London, United Kingdom

REVIEWED BY
Camilla Russo,
University of Naples Federico II, Italy
Giuseppe Pontillo,
University of Naples Federico II, Italy
Martina Di Stasi,
University of Naples Federico II, Italy

*CORRESPONDENCE
Bibiana Bielekova
Bibi.Bielekova@nih.gov

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Neuroradiology, a section of the journal Frontiers in Radiology

RECEIVED 23 August 2022
ACCEPTED 24 October 2022
PUBLISHED 11 November 2022

CITATION
Kelly E, Varosanec M, Kosa P, Prchkovska V, Moreno-Dominguez D and Bielekova B (2022) Machine learning-optimized Combinatorial MRI scale (COMRISv2) correlates highly with cognitive and physical disability scales in Multiple Sclerosis patients.
Front. Radio 2:1026442.
doi: 10.3389/fradi.2022.1026442

Machine learning-optimized Combinatorial MRI scale (COMRISv2) correlates highly with cognitive and physical disability scales in Multiple Sclerosis patients

Erin Kelly[1†], Mihael Varosanec[1†], Peter Kosa[1], Vesna Prchkovska[2], David Moreno-Dominguez[2] and Bibiana Bielekova[1*]

[1]Neuroimmunological Diseases Section, Laboratory of Clinical Immunology and Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, United States, [2]QMENTA, Boston, MA, United States

Composite MRI scales of central nervous system tissue destruction correlate stronger with clinical outcomes than their individual components in multiple sclerosis (MS) patients. Using machine learning (ML), we previously developed Combinatorial MRI scale (COMRISv1) solely from semi-quantitative (semi-qMRI) biomarkers. Here, we asked how much better COMRISv2 might become with the inclusion of quantitative (qMRI) volumetric features and employment of more powerful ML algorithm. The prospectively acquired MS patients, divided into training ($n = 172$) and validation ($n = 83$) cohorts underwent brain MRI imaging and clinical evaluation. Neurological examination was transcribed to NeurEx™ App that automatically computes disability scales. qMRI features were computed by lesion-TOADS algorithm. Modified random forest pipeline selected biomarkers for optimal model(s) in the training cohort. COMRISv2 models validated moderate correlation with cognitive disability [Spearman Rho = 0.674; Lin's concordance coefficient (CCC) = 0.458; $p < 0.001$] and strong correlations with physical disability (Spearman Rho = 0.830–0.852; CCC = 0.789–0.823; $p < 0.001$). The NeurEx led to the strongest COMRISv2 model. Addition of qMRI features enhanced performance only of cognitive disability model, likely because semi-qMRI biomarkers measure infratentorial injury with greater accuracy. COMRISv2 models predict most granular clinical scales in MS with remarkable criterion validity, expanding scientific utilization of cohorts with missing clinical data.

KEYWORDS

machine learning (ML), multiple sclerosis, MRI biomarkers, disability outcomes, predictive models

Introduction

Structural imaging of the central nervous system (CNS) by magnetic resonance (MRI) plays central role in diagnosing multiple sclerosis (MS) and evaluating efficacy of treatments. Nevertheless, the correlations between any MRI biomarker and clinical disability measures are only mild to moderate.

This is explainable by following shortcomings of both clinical scales and MRI biomarkers: A. Reliability: this includes technical aspects of the measurement such as test-retest variability, variability between different raters, different scanners or different analysis methods; and B. Criterion validity: this refers to how each measurement reflects true CNS tissue damage.

While original description of most expert-derived clinical scales missed test-retest reliability (e.g., Expanded Disability Status Scale [EDSS] (1)), the clinical trials identified "transient worsening and improvements" in approximately 20% of subjects (2), likely representing a measurement noise. We developed NeurEx™ App that eliminates part of the noise by algorithmically codified translation of a documented neurological examination into four clinical scales. Although the concordance correlation coefficient (CCC, reflects concordance of two ratings) of neuroimmunology scales between two clinicians transcribing the same documented neurological examination was excellent (i.e., ranging 0.943–0.968; p-value $< 1 \times 10^{-7}$), the difference for a single exam represented up to 3 EDSS points. By replacing one clinician with the NeurEx™ App that always provides only one rating per scale for a given documented exam, we increased inter-rater reliability to a maximum difference of 1.5 EDSS points (CCC 0.968–0.987; p-value $<1 \times 10^{-7}$). Of course, NeurEx™ can't eliminate noise stemming from variances in the performance of neurological examination by different clinicians and this likely represents the greater source of noise.

Even more pressing limitation of traditional clinical scales is their sensitivity and construct/criterion validity. For example, natural history cohorts show that on average an MS patient progresses by 1 EDSS point every 10 years (3, 4). Clearly, many axons demyelinate, and oligodendrocytes/neurons die during that time and this ongoing CNS tissue destruction is not captured by EDSS. Additionally, our ability to reliably quantify complex cognitive functions is extremely limited. Consequently, cognitive functions have been severely underrepresented in traditional disability scales. A creative attempt to remedy these limitations was MS functional composite (MSFC), an expert-derived composite scale of three functional tests reflecting ambulation, fine finger movements and memory/processing speed (5). While the concept of MSFC was outstanding, one of the selected components, Paced Auditory Serial Addition Test (PASAT3) proved suboptimal, suffering from high test-retest variability and a learning effect. This limitation was confirmed by developing Combinatorial, weight-adjusted Disability Score (CombiWISE; continuous scale from 0 to 100) (6) using data-driven approach to select contributing features and their optimal "weights"; and neither PASAT3 nor alternative cognitive test Symbol Digit Modalities Test [SDMT (7)] were selected by this model. Even though CombiWISE correlated strongly with EDSS in an independent cohort and measured significant disability progression over 6–12 months, this granular clinical scale still lacks sensitivity to measure destruction of individual axons/neurons and oligodendrocytes, likely happening in MS daily.

As the insensitivity of clinical scales to underlying cellular events is unsurmountable, MRI-based structural imaging and quantification of cellular substructures using advanced imaging methods such as magnetization transfer imaging (MTR) or diffusion tensor imaging (DTI) raised hopes for objective measurements of CNS tissue destruction. Unfortunately, MRI biomarkers proved to have their own limitations. MRI infers CNS structure from the signal decay of energized hydrogen protons, which is dependent on the technical aspects of specific MRI machine and acquisition protocols, on complex post-processing algorithms but also on transient biological processes such as subjects' hydration, use of alcohol or pharmaceutical agents (8). Consequently, test-retest variability of MRI biomarkers is high in comparison to measured change, leading to poor signal-to-noise ratio (SNR). The notable exception are semi-quantitative MRI features (semi-qMRI) such as number of contrast-enhancing lesions or number of (new) T2 lesions formed in different CNS compartments (6, 9, 10), which have excellent SNR.

Additionally, the criterion validity of any single MRI biomarker is problematic as all capture only some aspects of MS-related CNS tissue destruction and do so with restricted specificity. To surpass this limitation, several groups explored combinations of MRI features, akin to composite clinical scales (11–17). All published combinatorial MRI models outperformed each contributing MRI biomarker in correlation with clinical outcomes, validating this concept. Like in combinatorial clinical scales, most groups selected contributing MRI features based on expert opinions (18–21).

We took data driven approach to develop COMRISv1 (Combinatorial MRI scale, version 1), where both selection of contributing features and their weights in the final model were derived from unbiased machine learning (ML) approach (22). COMRISv1 was derived from semi-qMRI features only. This led to high SNR, while, inevitably, sacrificing sensitivity. Despite this, when tested in the independent validation cohort, COMRISv1 models achieved the highest correlations with physical (i.e., EDSS; Rho = 0.7, p-value < 0.001, $n = 114$) and cognitive (i.e., SDMT; Rho = 0.5, p-value < 0.0001, $n = 92$) disability among all published combinatorial MRI scales for MS. Nevertheless, we wondered, and this paper answers,

whether incorporating volumetric qMRI features and using more powerful ML models would strengthen performance of COMRISv2.

Materials and methods

Subjects and regulatory approvals

All subjects were prospectively recruited to the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) natural history protocol "Comprehensive Multimodal Analysis of patients with Neuroimmunological Diseases of the CNS"; Clinicaltrials.gov identifier NCT00794352. The study was approved by NIAID scientific review and by the NIH Institutional Review Board. All subjects provided written informed consent. **Supplementary Table S1** contains demographic and clinical data on all subjects.

Collection and computation of clinical scales

All participants underwent a comprehensive diagnostic process, including neurological examination transcribed to iPad-based App NeurEx™, which automatically calculates four clinical scales, including Expanded Disability Status Scale (EDSS; ordinal scale from 0 to 10) and Scripps Neurological Rating Scale (SNRS, continuous scale from 100 to 0) and streams data to Neuroimmunological Diseases Section (NDS) research database hosted on secured server. Another set of investigators, blinded to NeurEx data collected timed 25-foot walk (25FW), 9-hole peg test (9HPT) and SDMT and inputted these to the NDS database. The database automatically integrates data to calculate CombiWISE. NDS database has also user-defined privileges that blind the clinicians and investigators collecting clinical and functional data to qMRI and semi-qMRI data. MS diagnosis was based on 2010 McDonald diagnostic criteria (23) and, after 2017, based on its 2017 modifications (24).

Collection and computation of MRI biomarkers

Brain MRIs were performed on Signa – (1.5 T and 3 T, General Electric, Milwaukee, WI) and Skyra – (3 T, Siemens, Malvern, PA) units using 16 – and 32 – channel imaging coils with previously-described scanning protocols (22). Our brain MRI sagittal and axial cuts extend distally to C5 level, allowing determination of semi-qMRI biomarkers of medulla/upper spinal cord (SC) atrophy and lesion load.

The semi-qMRI data were acquired by consensus of MS-trained clinicians during weekly clinical care meetings. The rating of semi-qMRI features was previously extensively described (22) and its codification was integrated to NDS research database.

To acquire qMRI data, T1-magnetization-prepared rapid gradient-echo (MPRAGE) or fast spoiled gradient echo (FSPGR) and T2-weighted three-dimensional fluid attenuation inversion recovery (3D FLAIR) sequences, ideally with 1 mm^3 isotropic resolution, underwent a five-step pre-processing: (1) de-identification through the elimination of PHI-containing DICOM headers, (2) DICOM to NIFTI transformation, (3) 6-dof alignment to MNI template orientation, using *ANTS* package (25) to first register the T1 image to the 152 MNI template (26) and then co-register the T2 image to the aligned T1 image, (4) SkullStripping the T1 image using *ROBEX* (27) and using the same stripping mask to SkullStrip the co-registered T2 image, and (5) correct bias fields in the T1 image using the N4 algorithm from ANTS (28).

The volumetric data of different CNS structures were then computed by the LesionTOADS algorithm (29) implemented in a cloud based medical image-processing platform, QMENTA as part of collaborative project (https://catalog.qmenta.com/tool/lesion-toads-workflow). LesionTOADS uses an atlas-based technique combining a topological and statistical likelihood atlas for computation of following 12 segmented CNS tissues: Cerebral white matter, Cerebellar white matter, Brainstem, Putamen, Thalamus, Caudate, Cortical gray matter, Cerebellar gray matter, Lesion Volume, Ventricular CSF and Sulcal CSF.

LesionTOADS results were downloaded from QMENTA server and manually quality checked by an investigator blinded to clinical and functional data (MV). 17.2% of scans where LesionTOADS segmentation algorithm masks were incorrectly aligned with targeted anatomical structures were excluded from analyses.

Development and optimization of COMRISv2 models

COMRISv2 models were constructed using random forest (RF) (30), a decision-tree-based supervised learning algorithm. A decision tree is a modeling approach that uses multiple features (i.e., different MRI-based CNS volumes) to predict an outcome (i.e., disability) by finding the optimal split (e.g., a specific volume) at each branchpoint in the tree. Tree-based models are prone to "overfit" the data. RF aggregates thousands decision trees and uses a random subset of variables for decision-making at each branchpoint to limit (but not eliminate) the overfit problem. Thus, to further optimize our models, we used the iterative process where the least important variable ranked by variable importance

function was removed and the RF was rebuilt repeatedly until the root mean square error of the model reached its lowest point. For a visual depiction of this process, see Jackson et al. (31) and **Supplementary Figure S1**. Default tuning parameters (ntree = 500 and mtry = number of variables/3) were used for all models to ensure fair inter-model comparisons.

Statistical analyses and implemented safeguards to prevent bias

The correlation between observed and predicted outcomes was assessed by Spearman correlation coefficient Rho. The coefficient of determination (R^2) measuring the proportion of variance of observed outcomes that is explained by the model prediction, as well as the *p*-value, were calculated from linear regression models. The reproducibility of predicted vs. observed outcomes was evaluated by Lin's concordance correlation coefficient (CCC). The univariate correlations between age, clinical, and MRI outcomes was evaluated using Spearman correlation, the *p*-value cut-off for significant observations was set to 0.001 to account for 31 pairwise comparisons. All statistical analyses were performed in RStudio Version 1.1.463.

The user-defined privileges in the NDS database assured blinding, while software codification of the algorithms for calculating different scales prevented bias in these calculations. Finally, all models were validated in an independent cohort that did not participate in the model development.

Results

COMRISv2 model of cognitive disability: SDMT

The COMRISv2 model optimized to predict SDMT score validated in an independent cohort (Rho = 0.674, *p*-value <0.001, $R^2 = 0.458$, CCC = 0.562) (**Figure 1A**). COMRISv2 SDMT model outperformed the COMRISv1 predictions (Rho = 0.497, *p*-value <0.001, $R^2 = 0.247$, CCC = 0.404) of SDMT score in the same cohort. Age and qMRI features ranked most important in the model, although several semi-qMRI features were also included (**Figure 1B**).

COMRISv2 models of physical disability: EDSS, SNRS, CombiWISE and NeurEx

COMRISv2 models were also constructed to predict physical disability as measured by four different scales: EDSS, SNRS, CombiWISE, and NeurEx. All models of physical disability performed stronger than the COMRISv2 model for cognitive disability (**Figure 1C**). The NeurEx scale performed the strongest (Rho = 0.852, *p* value <0.001, $R^2 = 0.707$, CCC = 0.823). Models of physical disability favor semi-qMRI biomarkers reflecting disease burden in the infratentorial compartment (**Figure 1B**).

Comparing added value of quantitative volumetric features to COMRISv2 models

While semi-qMRI features can be easily collected by any trained clinician or even non-clinical investigator with knowledge of brain/spinal cord anatomy, collection of qMRI features require more specialized skillset and much more resources. To facilitate decisions about resource allocation, we formally assessed value of semi-qMRI and qMRI features for predicting different disability outcomes.

Thus, COMRISv2 models for SDMT were constructed considering only qMRI measures or only semi-qMRI measures, both in presence and absence of age. Cognitive disability models considering only qMRI measures (Rho = 0.568, *p*-value <0.001, $R^2 = 0.363$, CCC = 0.497) performed slightly better than those considering only semi-qMRI measures (Rho = 0.544, *p*-value <0.001, $R^2 = 0.282$, CCC = 0.474), but none outperformed the original model that integrates qMRI, semi-qMRI, and age (Rho = 0.674, *p*-value <0.001, $R^2 = 0.458$, CCC = 0.562).

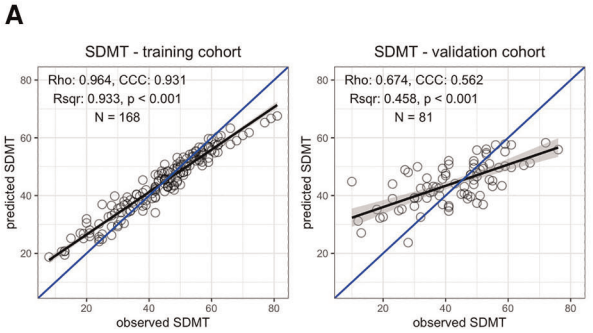
When we performed the same comparison in physical disability models (NeurEx and EDSS), addition of qMRI features did not improve model performance. NeurEx and EDSS models that considered only age and semi-qMRI features outperformed models that included qMRI features (**Figure 2**, **Supplementary Figure S2**).

Comparing feature selection between different COMRISv2 models with univariate correlations between MRI biomarkers and clinical outcomes

To facilitate interpretability of COMRIS models, we examined univariate correlations between all MRI features selected by at least one COMRIS model and all clinical outcomes plus age (**Figure 3**).

As would be expected, all clinical outcomes correlated moderately with age. qMRI outcomes related to CSF and Brain parenchymal fraction also correlated with age. From infratentorial structures, only cerebellar gray matter (GM) and semi-qMRI biomarker of brainstem and medulla/upper spinal cord (SC) atrophy correlated with age.

Most qMRI measures correlated with each other, except cerebellar GM, the only infratentorial qMRI biomarker selected by four out of five COMRIS models, which showed



B

	variable	variable importance					
		Σ all models	SDMT	EDSS	SNRS	CombiWISE	NeurEx
	Age	70.9	5.6	17.8	17.2	16.3	14.0
Semi-quantitative MRI features	Medulla atrophy	113.5	3.7	28.1	23.9	28.4	29.3
	Medulla lesion load	61.4	2.0	15.7	14.3	15.3	14.1
	Brainstem atrophy	52.8	2.6	8.4	15.5	13.3	13.0
	Cerebellum atrophy	7.4	2.4				5.0
	Brainstem lesion load	5.1	5.1				
	GM lesion load	4.0	4.0				
	Cerebellum lesion load	2.6	2.6				
	T1 black hole fraction	2.2	2.2				
	Brain atrophy severity						
	Type of brain atrophy						
	T2 lesion load						
	Juxtacortical lesion load						
	Mostly deep WM lesions						
	Prominent juxtacortical/cortical lesions						
	Prominent periventricular lesions						
	Prominent corpus callosum involvement						
	Thalamic or deep GM lesions						
	Superficila demyelination						
Quantitative MRI features	Cerebellum GM	28.3	4.9		9.2	8.9	5.3
	Supratentorial WM	27.5	4.3	6.1	10.1		7.1
	T2 lesions	27.4	12.1	5.5	9.8		
	Normal WM	27.2	10.1			10.3	6.8
	Brainstem	16.6	3.7	5.4		7.6	
	Ventricular CSF	14.8	8.8	5.9			
	Brain parenchymal fraction	11.6	4.6	7.1			
	Sulcal CSF	8.3	2.9				5.4
	Thalamus	5.3	5.3				
	Total CSF	4.1	4.1				
	Cerebellum WM	3.6	3.6				
	Cerebrum GM	2.7	2.7				
	Supratentorial GM	2.5	2.5				
	Caudate						
	Putamen						

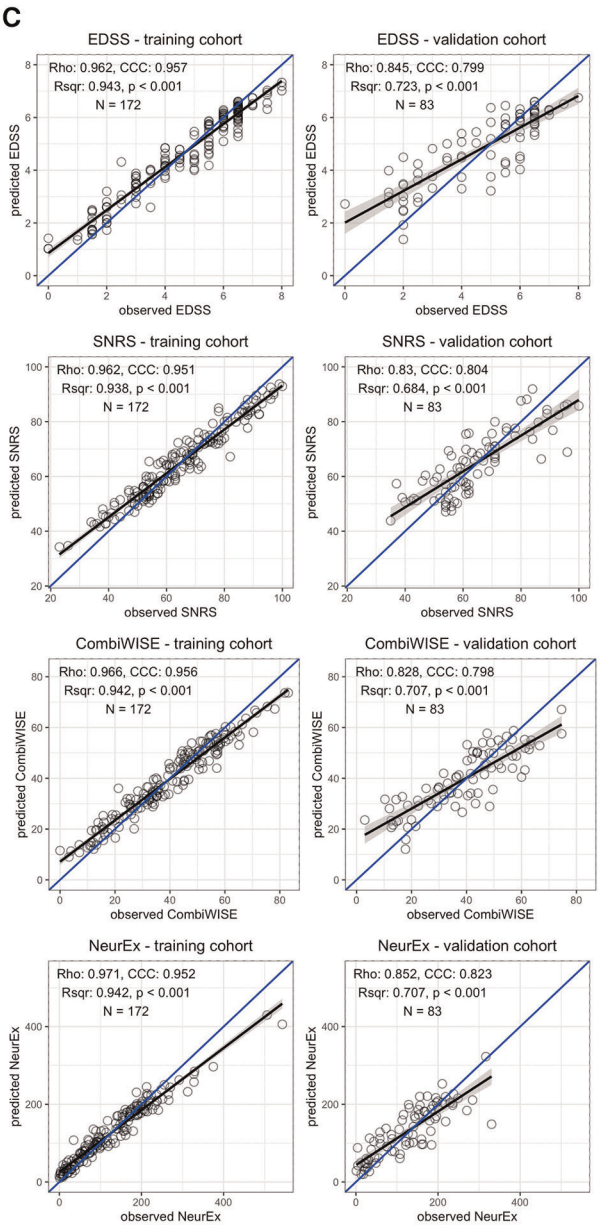


FIGURE 1
Random forest (RF)-based COMRISv2 models of clinical outcomes. (**A**) RF model of Symbol Digit Modalities Test (SDMT); (**B**) Age, 18 semi-quantitative, and 15 quantitative MRI features were used for RF modelling of the Expanded Disability Status Scale (EDSS), Scripps Neurological Rating scale (SNRS), Combinatorial Weight-adjusted Disability Score (CombiWISE), and digitalized neurological exam score (NeurEx). Variables selected by each model and their respective importance are highlighted here. Variables not selected by any model are depicted in gray. (**C**) RF models of physical disability measured by EDSS, SNRS, CombiWISE, and NeurEx. (**A**,**C**) The performance of each model was evaluated separately in the training cohort (left plot) and an independent validation cohort (right plot) by plotting observed values on the x-axis and model-predicted values on the y-axis. Spearman Rho, coefficient of determination (Rsqr), *p*-value (*p*), Lin's concordance correlation coefficient (CCC), and number of observations (**N**) are depicted. Black line represents a fitted linear model with the gray-shaded area corresponding to 95% confidence interval. The blue line represents 1:1 fit corresponding to 100% CCC.

only weak correlations with few outcomes. All semi-qMRI biomarkers correlated with each other, but the correlations were generally weak to moderate. Semi-qMRI features also correlated with qMRI features, except medulla/upper SC lesion load that correlated marginally with sulcal CSF. Although medulla/upper SC atrophy correlated with most of qMRI outcomes, these correlations were marginal with qMRI volumetric measures of telencephalon and strongest for brainstem and cerebellum GM volume. Thus, we conclude that qMRI and semi-qMRI measures provide

			SDMT					EDSS					NeurEx				
features included in modelling		Age	yes	yes		yes		yes	yes		yes		yes	yes		yes	
		semi-quantitative MRI	yes			yes	yes	yes			yes	yes	yes			yes	yes
		quantitative MRI	yes	yes	yes			yes	yes	yes			yes	yes	yes		
Model performance	training cohort	Spearman Rho	0.964	0.955	0.947	0.919	0.867	0.962	0.948	0.938	0.944	0.943	0.971	0.961	0.953	0.963	0.923
		CCC	0.931	0.918	0.911	0.866	0.815	0.957	0.921	0.872	0.935	0.932	0.952	0.901	0.886	0.946	0.914
		R^2	0.933	0.915	0.904	0.839	0.756	0.943	0.929	0.912	0.909	0.905	0.942	0.908	0.898	0.933	0.877
		adjusted R^2	0.933	0.915	0.903	0.838	0.755	0.942	0.929	0.912	0.909	0.904	0.942	0.908	0.897	0.933	0.876
		p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	validation cohort	Spearman Rho	**0.674**	0.568	0.556	0.544	0.486	0.845	0.613	0.465	**0.857**	0.829	0.852	0.630	0.461	**0.880**	0.848
		CCC	**0.562**	0.497	0.467	0.474	0.414	0.799	0.550	0.339	**0.844**	0.789	0.823	0.514	0.347	**0.866**	0.823
		R^2	**0.458**	0.363	0.324	0.282	0.220	0.723	0.437	0.205	**0.752**	0.665	0.707	0.361	0.183	**0.765**	0.688
		adjusted R^2	**0.451**	0.355	0.316	0.273	0.210	0.719	0.430	0.195	**0.749**	0.661	0.703	0.353	0.173	**0.762**	0.684
		p-value	**<0.001**	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	**<0.001**	<0.001	<0.001	<0.001	<0.001	**<0.001**	<0.001
Variable importance		Age	5.6	8.0		21.0		17.8	24.8		23.5		14.0	16.2		13.5	
	Semi-quantitative MRI features	Medulla atrophy	3.7			12.5	17.1	28.1			28.2	31.5	29.3			31.7	38.3
		Medulla lesion load	2.0					15.7			18.5	20.4	14.1			14.1	17.9
		Brainstem atrophy	2.6			13.5	16.6	8.4			9.9	10.5	13.0			16.3	18.6
		Cerebellum atrophy	2.4										5.0			6.4	9.3
		Brainstem lesion load	5.1			13.9	16.8				5.7	5.5				3.8	
		GM lesion load	4.0			12.7	16.3					6.6				4.4	7.2
		Cerebellum lesion load	2.6														
		T1 black hole fraction	2.2														
		Brain atrophy severity				14.8	18.1				8.4	8.3				6.5	8.8
		T2 lesion load				11.6	15.1					5.1					
		Juxtacortical lesion load									5.9	6.3				3.2	
		Type of brain atrophy										5.6					
		Mostly deep WM lesions															
		Prominent juxtacortical/cortical lesions															
		Prominent periventricular lesions															
		Prominent corpus callosum involvement															
		Thalamic or deep GM lesions															
		Superficila demyelination															
	Quantitative MRI features	Cerebellum GM	4.9	8.1	8.6				10.1	19.7			5.3	11.4	14.6		
		Supratentorial WM	4.3	6.3	8.0			6.1	10.0	19.6			7.1	14.1	15.7		
		T2 lesions	12.1	15.3	15.3			5.5	13.0	22.6				8.9	11.4		
		Normal WM	10.1	13.4	13.5								6.8	11.3	14.3		
		Brainstem	3.7	6.3	7.4			5.4	11.2	19.9				6.1	7.7		
		Ventricular CSF	8.8	11.7	12.1			5.9						7.4	9.2		
		Brain parenchymal fraction	4.6	6.0	7.9			7.1						5.9	9.4		
		Sulcal CSF	2.9	5.2									5.4	5.7	9.1		
		Thalamus	5.3	6.7	7.2				9.9								
		Total CSF	4.1	7.0	7.5				12.8					6.6			
		Cerebellum WM	3.6	6.1	6.9												
		Cerebrum GM	2.7		5.6												
		Supratentorial GM	2.5						8.1	18.3				6.4	8.7		
		Caudate															
		Putamen															

FIGURE 2
Evaluation of added value of quantitative MRI features. The original random forest (RF) models used age, semi-quantitative, and quantitative MRI features as an input. We tested how the RF models of Symbol Digit Modalities Test (SDMT), Expanded Disability Status Scale (EDSS), and digitalized neurological exam score (NeurEx) would perform if only semi-quantitative or only quantitative MRI features (with or without age) would perform. The performance of each model was evaluated separately in the training and an independent validation cohort by calculating Spearman Rho, Lin's concordance correlation coefficient (CCC), coefficient of determination (R^2), adjusted coefficient of determination (adjusted R^2), and *p*-value. The best performing models are highlighted by red rectangles. Variables selected by each model and their respective importance is also depicted.

mostly complimentary information, with qMRI outcomes better reflecting telencephalon tissue damage and semi-qMRI outcomes better capturing infratentorial tissue damage.

All clinical outcomes correlated with each other, with SDMT exhibiting only moderate correlations, while other clinical scales correlated strongly. The difference between SDMT and all remaining clinical outcomes was also evident from correlations with MRI outcomes: qMRI biomarkers correlated stronger with SDMT (cognitive disability) in comparison to clinical outcomes that capture predominantly physical disability. In contrast, semi-qMRI biomarkers, especially medulla/upper SC atrophy and lesion load, followed by brainstem atrophy, correlated with non-SDMT clinical outcomes and these correlations were overall stronger than correlations of qMRI measures with SDMT. All qMRI biomarkers correlated with physical disability outcomes (i.e., EDSS, SNRS, CombiWISE and NeurEx) weaker than age, whereas many qMRI outcomes outperformed age in correlation with SDMT.

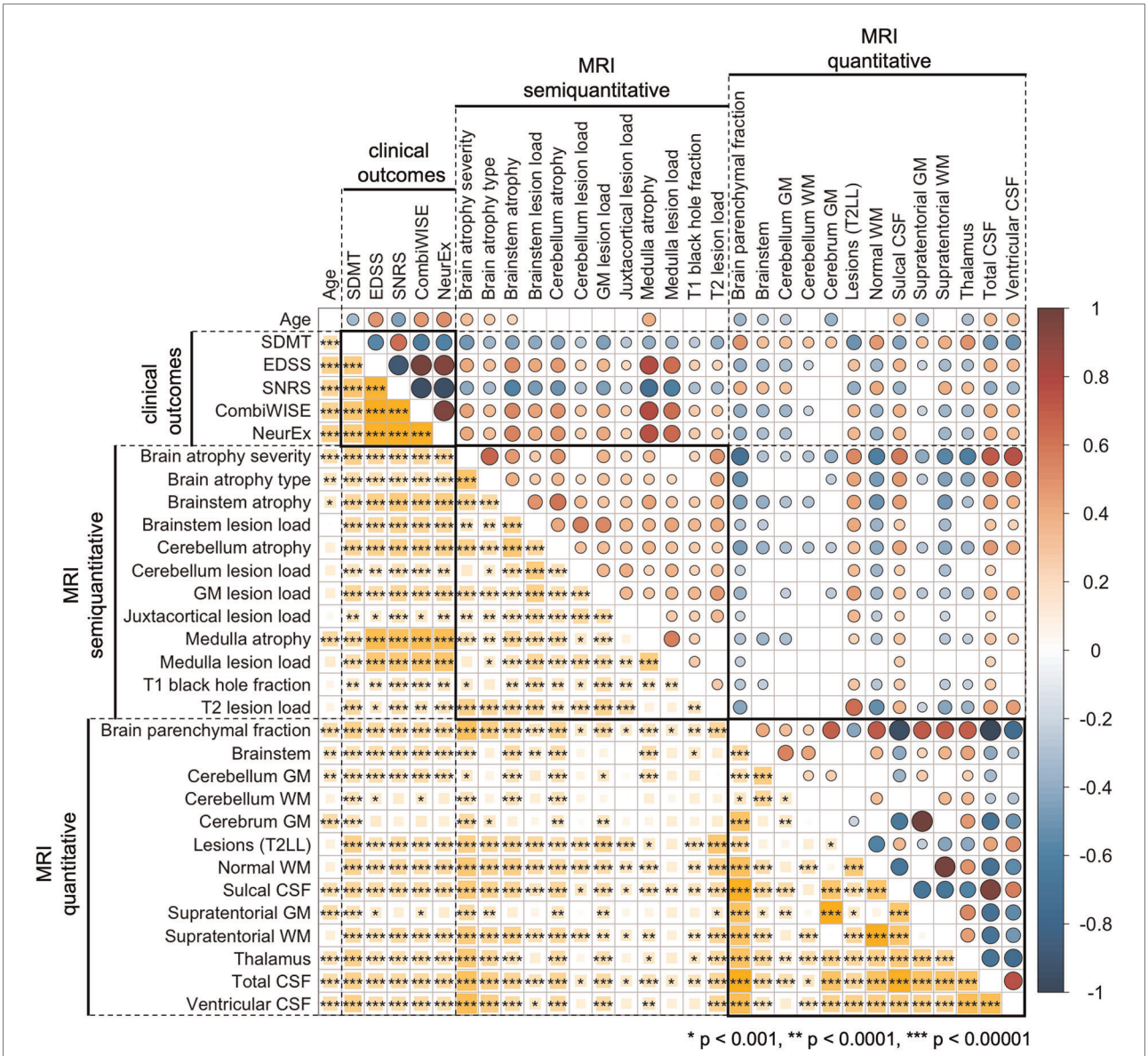


FIGURE 3
Univariate correlations between age, clinical outcomes, semiquantitative, and quantitative MRI features. Correlations were evaluated using Spearman Rho, with positive correlation in shades of red and negative correlations in shades of blue. The size of the circles above the diagonal and the size of the yellow squares below the diagonal corresponds to the absolute value of Rho. Only correlations with *p*-value below 0.001 were considered significant (accounting for 31 comparisons performed). The number of stars corresponds to the level of statistical significance.

Discussion

With the technological advances that allow reliable measurements of genetic, transcriptomic and proteomic biomarkers in hundreds of patients, the data scientists are realizing that the most limiting obstacle in translating these "omics" data into clinically translatable insights are, surprisingly, poor quality clinical and imaging data. This sentiment is epitomized in the recent review: "It is amazing how bad the standard data sets in the medical domain are (noisy, sparse, wrong, biased, etc)." (32) Employing unbiased, data-driven approaches to develop more accurate tools for measuring neurological disability and CNS tissue damage and validating both their criterion validity and reproducibility is the way to transcend this conundrum.

This paper demonstrates the power of ML approach to assemble semi-qMRI and qMRI brain imaging biomarkers into combinatorial models (COMRISv2) that reliably predict neurological disability in MS patients. Compared to previously published composite MRI scales, our study has following

strengths: (1) The MRI features and their weights are selected using unbiased, data-driven approach; (2) We studied a moderately large cohort of MS patients, with good representation of subjects with progressive MS; (3) COMRISv2 tested both semi-qMRI and qMRI volumetric data; (4) In addition to EDSS, we modeled COMRISv2 predictions of physical disability against SNRS, highly granular CombiWISE and NeurEx scales, and predictions of cognitive disability against SDMT; (5) Our models are validated in the independent cohort of MS patients that did not contribute to the development or optimization of the model(s).

We also recognize the following limitations of current study: (1) Although our original COMRISv1 computation is publicly available, including detailed guideline for semi-qMRI ratings (22), and we observed that adherence to those guidelines leads to mean interrater variability less than 10%, up till now no external group attempted to reproduce our data. This causes uncertainty whether other investigators could achieve analogous reproducibility of COMRIS models; (2) We did not test qMRI measures of atrophy or T2LL in the medulla/upper SC, as Lesion-TOADS algorithm does not provide these outcomes and also because we lacked dedicated SC MRI; (3) Our study did not include qMRI biomarkers derived from advanced imaging methods such as MTR or DTI.

While we can't influence the first limitation, we can address the effect of subsequent two limitations by literature review: First, high quality volumetric SC data require dedicated SC imaging, not available in our patients. Second, published observations suggest that addition of qMRI cervical SC biomarkers would have limited effect on COMRISv2 performance: e.g., the second iteration of Magnetic Resonance Disease Severity Scale (MRDSS2) (21) demonstrated that addition of upper cervical SC area to MRDSS1 model (which consisted only of brain qMRI features) increased correlation with EDSS from 0.25 to 0.33 ($p = 0.013$). Both COMRISv1 and COMRISv2 models (using only semi-qMRI features) validated much stronger correlations with EDSS (i.e., Rho = 0.857, $p < 0.001$). Analogously, meta-analysis (21 studies/1,933 participants) of dedicated 3 T SC imaging showed moderate correlation of cervical SC atrophy with EDSS (Rho = −0.42; $p < 0.0001$) (33). This likely represents over-estimation, as included studies with small number of participants showed invariably larger correlations. It has been convincingly shown that small studies over-estimate effect size (34). Correspondingly, large study (n = 1,249) published after the aforementioned meta-analysis measured Rho −0.315 ($p < 0.01$) for correlation of cervical SC volume with EDSS (35). These are analogous or smaller univariate correlations as those we observed in COMRIS models between EDSS and two highest ranking semi-qMRI biomarkers: medulla/cervical SC atrophy and T2LL [**Figure 3** and (22)].

Based on the limited value of volumetric qMRI compared to semi-qMRI biomarkers in COMRISv2 models of physical disability, we do not expect that incorporating MTR or DTI data could meaningfully enhance correlations with clinical outcomes for several reasons: (1) These advanced imaging biomarkers have even poorer SNR than volumetric qMRI measures (6, 10); and (2) COMRISv2 optimized for CombiWISE or NeurEx already explains close to 70% of physical disability variance in the independent validation cohort, which is exceptionally good performance.

To put our results into perspective, we performed a meta-analysis of 302 studies describing ML models of MS disability and severity outcomes (36), including 40 studies that modeled EDSS as an ordinal scale. Only half of those reported effect sizes. The meta-analysis evaluated published studies based on seven criteria (e.g., blinding, outlier removal, explanation of missingness, adjustment for confounders, adjustment for multiple comparison, presence of controls, and validation) and identified significant negative correlation between effects size and number of criteria fulfilled. An independent validation cohort, used in our study, that is essential to understand the true predictive power of composite construct on patients whose data did not contribute to model development, was missing in all published MRI studies predicting EDSS. Only one study out of 20 showed cross-validation results for EDSS models, achieving R^2 of 0.16–0.19 (37). In comparison, our optimized EDSS model explains 75% of variance in the independent validation cohort. Similar observation was made for MRI biomarker-based models of SDMT – 5 out of 12 studies reported effect sizes as R^2 ranging from 0.3 to 0.62 in the training cohort, compared to our optimized SDMT model that reaches R^2 of 0.933 in the training and 0.46 in the independent validation cohort. Presented data, congruent with most independent validation studies published, show unequivocally that training cohort results always over-estimate true strength of the relationships. Thus, we conclude that COMRISv2 models achieve the highest effect sizes in predicting clinical disability outcomes among published studies.

The limitation of the criterion validity of simple volumetric qMRI biomarkers we mentioned in the introduction is exemplified in highly informative post-mortem imaging pathological assessment, which showed that SC atrophy (19%–24%) strongly under-estimates axonal loss (57%–62%) in MS (38). Because these imaging data were postmortem, they were not affected by motion artifacts and signal averaging which would further decrease the strength of relationship between imaging biomarker and histologically measured CNS tissue destruction. Therefore, at very best technical imaging conditions the criterion validity of volumetric qMRI biomarkers of SC is limited.

Nevertheless, qMRI biomarkers, especially when measuring large telencephalon structures have validated relationship to CNS tissue destruction in MS: (1) Brain atrophy is higher in MS compared to HV; (2) It correlates with disability in large cohorts and (3) It predicts disability progression in long longitudinal studies (39–41). Consequently, if we could measure volume of all CNS structures with high accuracy, qMRI biomarkers would likely outperform semi-qMRI

biomarkers in all models, as we observed for SDMT version of COMRISv2.

Unfortunately, the test-retest variance ("noise") of qMRI outcomes increases inversely to the size and contrast of the structure measured and the required scanning time. Motion artifacts from skeletal muscles, heartbeats, breathing and cerebrospinal fluid pulsation disadvantage qMRI biomarkers from infratentorial structures compared to large telencephalic structures. Thus, volumetric biomarkers derived from small infratentorial structures with low MRI contrast from neighboring tissues, that need long acquisition times will have high "noise" (and low SNR) (6). This explains why infratentorial semi-qMRI biomarkers, while theoretically less sensitive, outperformed infratentorial qMRI features (**Figures 1B**, **2**): because they are measured with higher SNR. Research advances to limit measurement noise of infratentorial qMRI biomarkers may have greater clinical value than development of imaging methods that require longer scanning times and increased complexity of mathematical/physical data manipulations to produce quantitative output.

In conclusion, this study demonstrates excellent criterion validity of measuring CNS tissue damage in MS by two different modalities: neurological examination and brain MRI. There is nothing in clinical data that makes them inevitably poor quality (i.e., noisy, sparse, wrong, biased) if we approach their collection and their aggregation into sensitive and accurate scales with the same scientific rigor used to optimize collection and quantification of omics data. The observations that novel scales of neurological disability with much broader dynamic range than EDSS (i.e., total of 20 possible disability progression steps in the ordinal EDSS scale, vs. practically unlimited range of CombiWISE [0–100 continuous scale] and NeurEx [0 to theoretical maximum of 1,349] values) validated comparably, or even outperformed EDSS demonstrates that implementing data-driven approaches to development of new clinical scales allows increasing sensitivity without limiting their accuracy. The CCC of 0.866 between semi-qMRI features-derived COMRISv2 and NeurEx in the independent validation cohort indicates that scientists have at their fingertips a reliable inexpensive tool that can predict the most granular scale of neurological disability we currently have in MS research.

Data availability statement

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by NIAID scientific review and by the NIH Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

BB: designed the project and guided all aspects of data collection and analyses. VP and DMD: implemented LesionTOADS algorithm into QMENTA platform. MV: performed quality control of the volumetric MRI data. PK: managed research database, provided clean datasets for analyses, and generated figures. EK: performed statistical analyses and generated the machine learning models. EK, MV, PK and BB: wrote the first draft of the paper and all authors reviewed and edited the draft for intellectual content. All authors contributed to the article and approved the submitted version.

Funding

Funding for this study was provided by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health.

Acknowledgments

We would like to thank our clinical team, patients, and caregivers at the Neuroimmunological Diseases Section for partnering with us in this project.

Conflict of interest

VP and DMD were employed by QMENTA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fradi.2022.1026442/full#supplementary-material.

References

1. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*. (1983) 33(11):1444–52. doi: 10.1212/WNL.33.11.1444

2. Ebers GC, Heigenhauser L, Daumer M, Lederer C, Noseworthy JH. Disability as an outcome in MS clinical trials. *Neurology*. (2008) 71(9):624–31. doi: 10.1212/01.wnl.0000313034.46883.16

3. Confavreux C, Vukusic S, Moreau T, Adeleine P. Relapses and progression of disability in multiple sclerosis. *N Engl J Med*. (2000) 343(20):1430–8. doi: 10.1056/NEJM200011163432001

4. Confavreux C, Vukusic S. The clinical course of multiple sclerosis. *Handb Clin Neurol*. (2014) 122:343–69. doi: 10.1016/B978-0-444-52001-2.00014-5

5. Fischer JS, Rudick RA, Cutter GR, Reingold SC. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. *Mult Scler*. (1999) 5(4):244–50. doi: 10.1177/135245859900500409

6. Kosa P, Ghazali D, Tanigawa M, Barbour C, Cortese I, Kelley W, et al. Development of a sensitive outcome for economical drug screening for progressive multiple sclerosis treatment. *Front Neurol*. (2016) 7:131. doi: 10.3389/fneur.2016.00131

7. Benedict RH, DeLuca J, Phillips G, LaRocca N, Hudson LD, Rudick R, et al. Validity of the Symbol Digit Modalities Test as a cognition performance outcome measure for multiple sclerosis. *Mult Scler*. (2017) 23(5):721–33. doi: 10.1177/1352458517690821

8. Weinberger DR, Radulescu E. Finding the elusive psychiatric "lesion" with 21st-century neuroanatomy: a note of caution. *Am J Psychiatry*. (2016) 173(1):27–33. doi: 10.1176/appi.ajp.2015.15060753

9. Borges IT, Shea CD, Ohayon J, Jones BC, Stone RD, Ostuni J, et al. The effect of daclizumab on brain atrophy in relapsing-remitting multiple sclerosis. *Mult Scler Relat Disord*. (2013) 2(2):133–40. doi: 10.1016/j.msard.2012.10.002

10. Oh J, Chen M, Cybulsky K, Suthiphosuwan S, Seyman E, Dewey B, et al. Five-year longitudinal changes in quantitative spinal cord MRI in multiple sclerosis. *Mult Scler*. (2020) 27(4):549–58. doi: 10.1177/1352458520923970

11. Filippi M, Paty DW, Kappos L, Barkhof F, Compston DA, Thompson AJ, et al. Correlations between changes in disability and T2-weighted brain MRI activity in multiple sclerosis: a follow-up study. *Neurology*. (1995) 45(2):255–60. doi: 10.1212/WNL.45.2.255

12. Mainero C, De Stefano N, Iannucci G, Sormani MP, Guidi L, Federico A, et al. Correlates of MS disability assessed in vivo using aggregates of MR quantities. *Neurology*. (2001) 56(10):1331–4. doi: 10.1212/WNL.56.10.1331

13. Riahi F, Zijdenbos A, Narayanan S, Arnold D, Francis G, Antel J, et al. Improved correlation between scores on the expanded disability status scale and cerebral lesion load in relapsing-remitting multiple sclerosis. Results of the application of new imaging methods. *Brain*. (1998) 121(Pt 7):1305–12. doi: 10.1093/brain/121.7.1305

14. Sailer M, Losseff NA, Wang L, Gawne-Cain ML, Thompson AJ, Miller DH. T1 lesion load and cerebral atrophy as a marker for clinical progression in patients with multiple sclerosis. A prospective 18 months follow-up study. *Eur J Neurol*. (2001) 8(1):37–42. doi: 10.1046/j.1468-1331.2001.00147.x

15. Wolinsky JS, Borresen TE, Dietrich DW, Wynn D, Sidi Y, Steinerman JR, et al. GLACIER: an open-label, randomized, multicenter study to assess the safety and tolerability of glatiramer acetate 40 mg three-times weekly versus 20 mg daily in patients with relapsing-remitting multiple sclerosis. *Mult Scler Relat Disord*. (2015) 4(4):370–6. doi: 10.1016/j.msard.2015.06.005

16. Wolinsky JS, Narayana PA, Johnson KP. United States open-label glatiramer acetate extension trial for relapsing multiple sclerosis: MRI and clinical correlates. Multiple Sclerosis Study Group and the MRI Analysis Center. *Mult Scler*. (2001) 7(1):33–41. doi: 10.1177/135245850100700107

17. Wolinsky JS, Narayana PA, Noseworthy JH, Lublin FD, Whitaker JN, Linde A, et al. Linomide in relapsing and secondary progressive MS: part II: MRI results. MRI Analysis Center of the University of Texas-Houston, Health Science Center, and the North American Linomide Investigators. *Neurology*. (2000) 54(9):1734–41. doi: 10.1212/WNL.54.9.1734

18. Bakshi R, Neema M, Healy BC, Liptak Z, Betensky RA, Buckle GJ, et al. Predicting clinical progression in multiple sclerosis with the magnetic resonance disease severity scale. *Arch Neurol*. (2008) 65(11):1449–53. doi: 10.1001/archneur.65.11.1449

19. Moodie J, Healy BC, Buckle GJ, Gauthier SA, Glanz BI, Arora A, et al. Magnetic resonance disease severity scale (MRDSS) for patients with multiple sclerosis: a longitudinal study. *J Neurol Sci*. (2012) 315(1-2):49–54. doi: 10.1016/j.jns.2011.11.040

20. Yousuf F, Kim G, Tauhid S, Glanz BI, Chu R, Tummala S, et al. The contribution of cortical lesions to a composite MRI scale of disease severity in multiple sclerosis. *Front Neurol*. (2016) 7:99. doi: 10.3389/fneur.2016.00099

21. Bakshi R, Neema M, Tauhid S, Healy BC, Glanz BI, Kim G, et al. An expanded composite scale of MRI-defined disease severity in multiple sclerosis: MRDSS2. *Neuroreport*. (2014) 25(14):1156–61. doi: 10.1097/WNR.0000000000000244

22. Kosa P, Komori M, Waters R, Wu T, Cortese I, Ohayon J, et al. Novel composite MRI scale correlates highly with disability in multiple sclerosis patients. *Mult Scler Relat Disord*. (2015) 4(6):526–35. doi: 10.1016/j.msard.2015.08.009

23. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann Neurol*. (2011) 69(2):292–302. doi: 10.1002/ana.22366

24. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol*. (2018) 17(2):162–73. doi: 10.1016/S1474-4422(17)30470-2

25. Avants BTN, Song G. Advanced normalization tools: V1.0. *Insight J*. (2009) 2(365):1–35. doi: 10.54294/uvnhin

26. Grabner G, Janke AL, Budge MM, Smith D, Pruessner J, Collins DL. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. *Med Image Comput Comput Assist Interv*. (2006) 9(Pt 2):58–66. doi: 10.1007/11866763_8

27. Iglesias JE, Liu CY, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans Med Imaging*. (2011) 30(9):1617–34. doi: 10.1109/TMI.2011.2138152

28. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. (2010) 29(6):1310–20. doi: 10.1109/TMI.2010.2046908

29. Shiee N, Bazin PL, Ozturk A, Reich DS, Calabresi PA, Pham DL. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *Neuroimage*. (2010) 49(2):1524–35. doi: 10.1016/j.neuroimage.2009.09.005

30. Breiman L. Random forests. *Mach Learn*. (2001) 45(1):5–32. doi: 10.1023/A:1010933404324

31. Jackson KC, Sun K, Barbour C, Hernandez D, Kosa P, Tanigawa M, et al. Genetic model of MS severity predicts future accumulation of disability. *Ann Hum Genet*. (2020) 84(1):1–10. doi: 10.1111/ahg.12342

32. Holzinger A, Haibe-Kains B, Jurisica I. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *Eur J Nucl Med Mol Imaging*. (2019) 46(13):2722–30. doi: 10.1007/s00259-019-04382-9

33. Song X, Li D, Qiu Z, Su S, Wu Y, Wang J, et al. Correlation between EDSS scores and cervical spinal cord atrophy at 3 T MRI in multiple sclerosis: a systematic review and meta-analysis. *Mult Scler Relat Disord*. (2020) 37:101426. doi: 10.1016/j.msard.2019.101426

34. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. (2008) 19(5):640–8. doi: 10.1097/EDE.0b013e31818131e7

35. Andelova M, Uher T, Krasensky J, Sobisek L, Kusova E, Srpova B, et al. Additive effect of spinal cord volume, diffuse and focal cord pathology on disability in multiple sclerosis. *Front Neurol*. (2019) 10:820. doi: 10.3389/fneur.2019.00820

36. Liu J, Kelly E, Bielekova B. Current status and future opportunities in modeling clinical characteristics of multiple sclerosis. *Front Neurol*. (2022) 13:884089. doi: 10.3389/fneur.2022.884089

37. Cordani C, Hidalgo de la Cruz M, Meani A, Valsasina P, Esposito F, Pagani E, et al. MRI correlates of clinical disability and hand-motor performance in multiple sclerosis phenotypes. *Mult Scler*. (2021) 27(8):1205–21. doi: 10.1177/1352458520958356

38. Petrova N, Carassiti D, Altmann DR, Baker D, Schmierer K. Axonal loss in the multiple sclerosis spinal cord revisited. *Brain Pathol*. (2018) 28(3):334–48. doi: 10.1111/bpa.12516

39. Chard DT, Griffin CM, Parker GJ, Kapoor R, Thompson AJ, Miller DH. Brain atrophy in clinically early relapsing-remitting multiple sclerosis. *Brain*. (2002) 125(Pt 2):327–37. doi: 10.1093/brain/awf025

40. Ge Y, Grossman RI, Udupa JK, Wei L, Mannon LJ, Polansky M, et al. Brain atrophy in relapsing-remitting multiple sclerosis and secondary progressive multiple sclerosis: longitudinal quantitative analysis. *Radiology*. (2000) 214(3):665–70. doi: 10.1148/radiology.214.3.r00mr30665

41. Jacobsen C, Hagemeier J, Myhr KM, Nyland H, Lode K, Bergsland N, et al. Brain atrophy and disability progression in multiple sclerosis patients: a 10-year follow-up study. *J Neurol Neurosurg Psychiatry*. (2014) 85(10):1109–15. doi: 10.1136/jnnp-2013-306906