Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

Assessing the role of volumetric brain information in multiple sclerosis progression

Andy A. Shen [a],[ID],[1], Aidan McLoughlin [b],[ID],[1], Zoe Vernon [a], Jonathan Lin [c],[ID], Richard A.D. Carano [d],[ID], Peter J. Bickel [a,b], Zhuang Song [d],[*], Haiyan Huang [a,b],[*],[**]

[a] *Department of Statistics, UC Berkeley, Berkeley, CA, USA*
[b] *Division of Biostatistics, UC Berkeley, Berkeley, CA, USA*
[c] *Department of Statistical Science, Duke University, Durham, NC, USA*
[d] *Analytics and Medical Imaging, Product Development, Genentech Inc., South San Francisco, CA, USA*

ARTICLE INFO

ABSTRACT

Multiple sclerosis is a chronic autoimmune disease that affects the central nervous system. Understanding multiple sclerosis progression and identifying the implicated brain structures is crucial for personalized treatment decisions. Deformation-based morphometry utilizes anatomical magnetic resonance imaging to quantitatively assess volumetric brain changes at the voxel level, providing insight into how each brain region contributes to clinical progression with regards to neurodegeneration. Utilizing such voxel-level data from a relapsing multiple sclerosis clinical trial, we extend a model-agnostic feature importance metric to identify a robust and predictive feature set that corresponds to clinical progression. These features correspond to brain regions that are clinically meaningful in MS disease research, demonstrating their scientific relevance. When used to predict progression using classical survival models and 3D convolutional neural networks, the identified regions led to the best-performing models, demonstrating their prognostic strength. We also find that these features generalize well to other definitions of clinical progression and can compensate for the omission of highly prognostic clinical features, underscoring the predictive power and clinical relevance of deformation-based morphometry as a regional identification tool.

1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease that affects the central nervous system (CNS) in young adults [1]. Magnetic resonance imaging (MRI) of the brain is a key tool for diagnosing and monitoring MS progression. Different types of MRI data possess overlapping yet complementary properties and applications. For instance, brain volume measurements derived from T1-weighted MRI provide insights into brain atrophy, which is associated with disability and disease progression [2–5]. Identifying specific brain regions with atrophy and volume changes that are linked to disease progression enhances patient

monitoring and facilitates the evaluation of treatment efficacy [6]. Early efforts to assess volume changes in MS consisted of using voxel-based morphometry (VBM), a method designated for volume assessment of gray matter regions [7]. VBM demonstrated limited associations with clinical outcomes in MS studies [8] and a majority of VBM studies were cross-sectional and thus unable to track disease-related changes at the individual level. On the other hand, deformation-based morphometry (DBM) [9,10] provides voxel-level measurements of volume assessment across the entire brain, including gray matter, white matter, and the ventricles. We note that enlargement of the fluid-filled ventricles in the brain typically reflect either local or global brain tissue loss, which can-

* Z. Song and H. Huang are co-last, co-corresponding authors.
** Corresponding author at: 367 Evans Hall, Berkeley, CA 94720, USA.
E-mail addresses: song.zhuang@gene.com, hhuang@stat.berkeley.edu (Z. Song).
[1] A. Shen and A. McLoughlin contributed equally to this work.

not be captured by VBM due to its focus on gray matter regions. While VBM often suffers from tissue segmentation errors [10,11], DBM only relies on image registration, making it a more robust tool to assess voxel-based volume differences.

MRI data has proven instrumental in predicting disease progression and plays a critical role in personalizing patient treatment decisions [12,13]. Despite recent progress in using T1-weighted MRI data to predict MS progression [14,15,4,16], several open questions remain regarding how data and methodological choices may affect model reliability. For example, recent work leveraging random survival forests achieved high accuracy in predicting clinical progression over a 10-year period based on changes in T1-weighted MRI-derived features within the first 2 years [17]. However, baseline T1-weighted MRI data alone demonstrated relatively limited predictive power for 2-year progression [18]. Similarly, it was shown that, while voxel clusters derived from independent components analysis (ICA) applied to 2-year volume change data were predictive of 10-year progression, the voxel cluster values at baseline were similar between progressors and non-progressors [19]. Some studies have also noted unclear predictive value of T1-weighted MRI, which can vary between studies and can depend on outcome definitions [18,20]. These issues also persist in deep learning frameworks [21,22]. It also remains unclear whether T1-weighted MRI data provides complementary information for progression prediction beyond clinician-measured variables of disease state, some of which are used to define common progression endpoints themselves [23,24]. Ultimately, there is a lack of literature on detecting baseline volume measurements from T1-weighted brain MRI that are predictive of clinical progression within two years. For broader reviews of data-driven approaches to clinical modeling of MS, we refer readers to [14,15,4,16,25].

In this study, we utilize baseline deformation-based morphometry derived from T1-weighted MRI of a Phase III MS clinical trial [26,6] to investigate the impact of various methodological choices, particularly those used to identify predictive DBM features, on the reliability of conclusions drawn from this dataset. We first employ a state-of-the-art, model-agnostic feature importance algorithm, adapted using leave-one-covariate-out methods [27] and minipatch ensembles (LOCO-MP) [28], to identify brain regions from low-signal T1-weighted MRI baseline data that are predictive of MS progression. Broadly speaking, LOCO-MP computes feature importance by aggregating predictions from models trained on multiple small, randomly selected subsets of the data (minipatches). We observe that LOCO-MP combined with a survival random forest model identifies features in a more consistent and stable manner than survival random forests alone. This is because LOCO-MP leverages learning from small feature subsets to isolate the effects of highly correlated regions while other methods often yield inconsistent results when the data are highly correlated and high-dimensional [29].

Next, we apply traditional survival models to examine how different outcome/endpoint definitions, in conjunction with feature selection by LOCO-MP, influence prediction power. Although our progression endpoint (CCDP24, discussed in Section 2.1) is highly censored within the two-year study period, it represents a well-defined and clinically meaningful composite marker of MS progression. To validate the DBM features identified using LOCO-MP, we assess their generalizability across alternative progression endpoints (S25FW, discussed in Section 2.1) to compare a single, objective measure of physical disability with the composite measure, the latter involving some degree of clinician discretion [23]. We also verify that our selected features generalize well across a separate patient cohort (see Appendix D).

By using DBM at baseline to identify patient volumetric abnormalities and by applying LOCO-MP to enforce sparsity in the DBM feature set, we achieve improved prediction performance compared to models that use whole-brain T1-weighted MRI data as input. Additionally, we compare the model performance of DBM features against models using conventional, non-DBM clinical features (such as age and sex), also measured at baseline. We introduce these conventional features in Section 2.1. Interestingly, models using selected MRI features alone perform

on par with those using only the baseline conventional measurements. Combined models incorporating both selected MRI features and conventional measures do not necessarily outperform individual models (those using either selected MRI features or conventional measures alone). Notably, removing the conventional features that define certain endpoints significantly reduces the predictive power of the conventional feature models, underscoring the strong predictive value of the selected MRI features.

We further analyze whether 3D convolutional neural networks (CNNs) using only the LOCO-MP regions perform better than full brain architectures. We develop a 3D CNN architecture, *Region CNN*, that accepts a set of atlas region tensors as input. While full brain MRI deep transformer models [30,31] are unstable and show low prediction accuracy, *Region CNN* using LOCO-MP identified features performs substantially better. This highlights the importance in filtering out unnecessary regions via LOCO-MP for extracting prognostic signal from CNNs for this cohort.

2. Materials and methods

2.1. Data and image preprocessing

The data used in this study are from a comparator arm of a phase 3 clinical trial of relapsing multiple sclerosis (OPERA I: NCT01247324), in which the patients were treated with interferon (IFN) β-1a (44 μg) three times per week throughout the 96-week treatment period. Our sample consisted of 350 patients. Details on patient selection, MRI acquisition, and clinical assessments are provided in the original report [26]. Briefly, patients were recruited with an age range of 18 to 55 years; McDonald criteria diagnosis of multiple sclerosis; Expanded Disability Status Scale (EDSS) score between 0 to 5.5; more than 2 documented clinical relapses within the previous 2 years or one clinical relapse within the year before screening; brain MRI evidence of multiple-sclerosis-related abnormalities. Conventional T1-weighted 3D spoiled gradient-recalled echo brain MRI was acquired at baseline, Weeks 24, 48 and 96 (repetition time = 28–30 ms, echo time = 5–11 ms, flip angle = 27–30 deg, 60 oblique axial slices of 1 mm in-plane resolution and 3 mm slice thickness). The primary clinical outcome was defined as a composite measure of disability progression which included three clinical assessments: EDSS, Timed 25-Foot Walk (T25FW, a measure of short distance walking speed), and 9-Hole Peg Test (9HPT, a measure of upper limb function) [32]. This composite measure was developed to capture a broader aspect of disability in patients with multiple sclerosis. The 24-week composite confirmed disability progression (CCDP24) was defined as progression on any one of the three components (EDSS, T25FW, or 9HPT): an increase of EDSS score from the baseline at least 1.0 point (or 0.5 points if the baseline EDSS score was larger than 5.5) or a 20% minimum threshold change for T25FW and 9HPT. We also consider a 24-week progression outcome based solely on the 25-foot walking score (S25FW) in Section 3.2.2. Table 1 summarizes the conventional features we use for prediction, grouped by those who experienced CCDP24 progression and those who were censored.

Since MS progression is influenced by many factors, our models consider a variety of demographic and clinical measurements as well as standard imaging metrics. Broadly speaking, we refer to these features as "conventional features" throughout the manuscript to distinguish them from DBM-based features. The conventional features serve as a comparison against the DBM features to assess which feature set performs more optimally. We record nine conventional features which include the patient's age, birth sex, years of onset, weight (in kg), total brain volume, total T2 lesion volume,[2] baseline 25-foot walking test score, baseline

---

[2] The T2-weighted MRI sequence was described by Elliot et al. (2019) [32]: axial 3 mm T2-weighted slices were acquired with 2D fast spin-echo, repetition time = 4000–6190 ms, echo time = 74–91 ms and echo train length =

Table 1
Baseline characteristics stratified by CCDP24 outcome (censored/progressed).

Variable	Censored (N = 269) Mean (SD) or n / N (%)	Progressed (N = 81) Mean (SD) or n / N (%)	p-value
Baseline age (years)	37 (9)	37 (10)	> 0.9
Years of onset	6.3 (6.2)	7.3 (6.0)	0.088
Baseline weight (kg)	76 (18)	74 (16)	0.4
Baseline 9-hole peg test score time (seconds)	23.1 (6.6)	24.2 (5.2)	0.033
Baseline 25-foot walking test score time (seconds)	7.0 (5.0)	5.9 (3.7)	0.049
Baseline EDSS	2.73 (1.28)	2.57 (1.25)	0.4
T2 lesion volume	8 (9)	13 (14)	0.019
Brain volume	1,501 (86)	1,480 (85)	0.066
Proportion female	185 / 269 (69%)	50 / 81 (62%)	0.2

Wilcoxon rank sum test for continuous variables; Pearson's Chi-squared test for categorical variable.
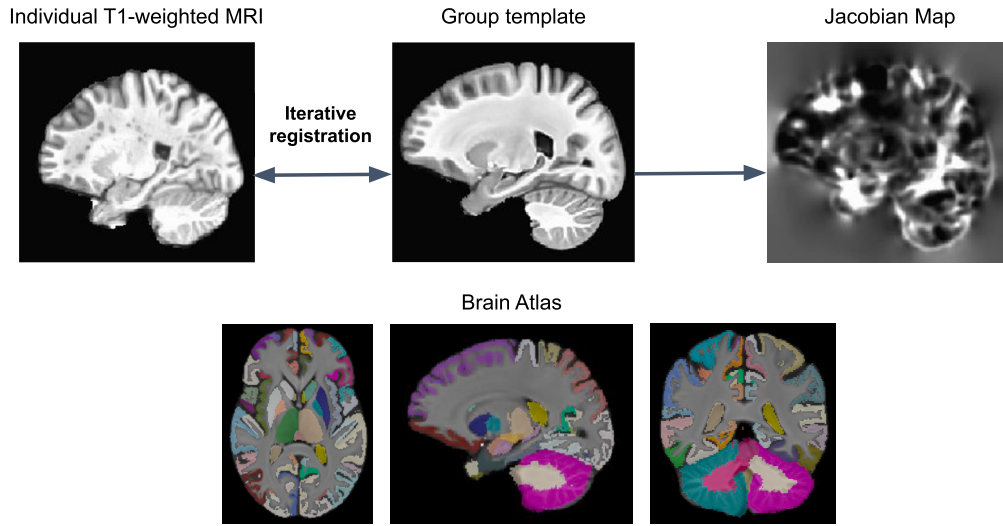

Fig. 1. Illustration of the deformation-based morphometry (DBM) analysis. For each T1-weighted brain MR image, the registration with the group template generates a Jacobian map over the whole brain, which is a voxel-level measure of volume relative to the group template. Regional labels on the brain atlas were mapped to the group template to extract statistical features of the Jacobian maps at the patient level.

9-hole peg test score, and baseline EDSS, which are generally thought to be related to MS disease progression [34].

Statistical features of regional brain volume were extracted from T1-weighted brain MRI, using a deformation-based morphometry (DBM) pipeline based on the diffeomorphic image registration of advanced normalization tools [35]. (See Fig. 1.) The Mindboggle atlas was used to identify individual brain regions [36]. In addition, prior probability images of six major brain regions based on the Mindboggle atlas, including CSF, cortical gray matter, deep gray matter, cortical white matter, the brainstem, and the cerebellum, were binarized at a probability threshold of 0.5. A population-specific group template of the T1-weighted brain MRI was constructed by advanced normalization tools with T1-weighted brain MRI images of 171 healthy adults (aged 20–59 years). Before being fed into the pipeline, each individual brain image was preprocessed with the following steps: 1) resampling to an isotropic resolution of 1 mm; 2) N4 bias correction [37]; and 3) denoising with a non-local algorithm [38]. Regional labels were mapped from the group template to individual images so that volume change of any predefined region could be measured.

As mentioned in the introduction, all of the conventional and DBM features we use for prediction are measured at baseline. The tabular DBM feature set consists of the median and standard deviation voxel values for each of the 51 atlas regions, resulting in 102 total features. Prior to performing feature selection and modeling, we filter out features with a variance below 0.01 across all patients, reducing the number of DBM features from 102 to 56. No other conventional features were excluded in this filtering step.

2.2. Feature importance using minipatch ensembles

Identifying features that genuinely contribute to prediction is crucial for model interpretability and guiding practical decision-making. Feature selection methods are used to filter out variables that are inconsequential to prediction, retaining those with the most predictive value. Since our dataset consists of high-dimensional and highly correlated features, traditional feature selection methods can potentially misattribute the importance of a feature when other correlated features are present. To offer a more reliable assessment of feature importance, we extend a feature importance method called *leave-one-covariate-out minipatch (LOCO-MP)* prediction [27,39,40,28] to survival analysis. LOCO-MP consists of taking a subsample of the data, called a "minipatch." The purple cells in Fig. 2 shows a simple example of what three minipatches may look like. A model is then trained on the minipatch and this process is repeated over many minipatches to ensure an even distribution of observations and features are sampled. After training on all minipatches, LOCO-MP determines feature importance for a certain target feature j by comparing the average prediction error across minipatches with and without feature j.

7–11. The T2 lesion volume was measured from T2-weighted MRI images using a proprietary, semi-automatic method. The initial automatic segmentation was performed using a Bayesian classifier [33] which was then verified and corrected by qualified readers.
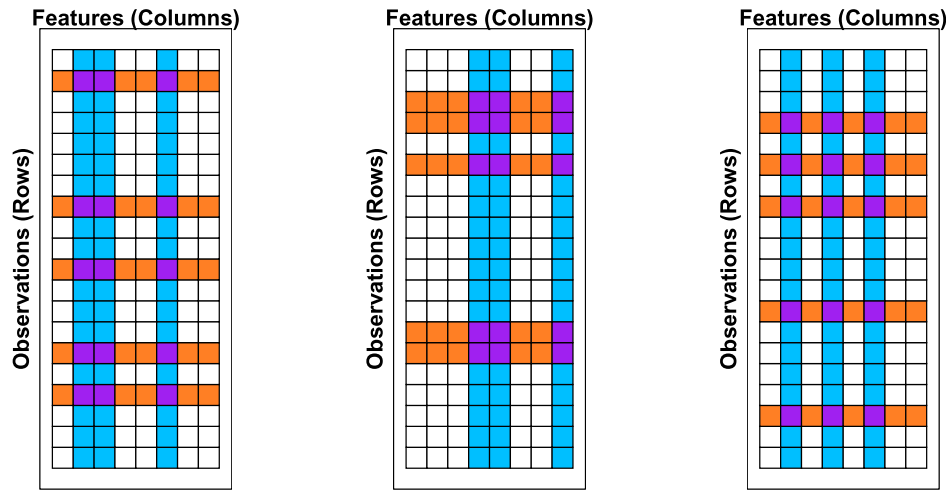
Fig. 2. General overview of LOCO-MP. Each plot demonstrates how a subset of observations (orange) and features (blue) are subsampled to form a "minipatch" (purple cells). A model is trained on the purple cells and predictions are generated using the unselected observations (blue cells). This process is repeated over numerous minipatches to ensure each feature and observation is selected enough times to perform inference. Feature importance for a target feature is assessed by evaluating performance on minipatches with and without that feature.

LOCO-MP is advantageous over other feature selection techniques due to its ability to provide asymptotic inferential guarantees [28]. Moreover, LOCO-MP can also account for dependencies across features, a common issue in high-dimensional settings like ours – Gan et al. (2023) [28] discuss that, by generating randomly subsampled features across minipatches, LOCO-MP ensures that the predictive value of each feature is not being diminished by other strongly correlated features, since groups of strongly correlated features will not always appear in the same minipatch. The full LOCO-MP procedure and its theoretical guarantees are discussed in Gan et al. (2023) [28].

The procedure is summarized graphically in Fig. 2 and outlined below for a target feature j:

1. **Subsample the data:** Subsample a set of observations (rows) and features (columns), denoted by purple cells. This forms a "minipatch."
2. **Train a model:** Fit a model to the minipatch and evaluate predictions using the remaining observations (blue cells).
3. Repeat steps 1 and 2 over K minipatches, where K is user-determined.
4. **Compute feature importance:** Compute $\bar{\Delta}_j$, the feature importance score of feature j (discussed in the following paragraphs).

Our outcome of interest, patient progression, is recorded as a time-to-event pair (T_i, C_i), where T_i is the event time and $C_i \in \{0, 1\}$ is an indicator for whether the patient's disease progression is observed during the study period ($C_i = 1$) or is right-censored ($C_i = 0$). Because our outcomes of interest are recorded in this time-to-event format, we extend LOCO-MP to survival analysis. This is done by predicting discrete instantaneous hazards from the trained models and computing an associated individual-level discrete hazard loss [41,42]. We refer the reader to Appendix A for a full technical description of LOCO-MP and the hazard loss function used to characterize prediction error in our survival models.

2.3. MS progression with classical survival models

To assess the predictive power of the DBM and conventional features, we apply random survival forests (RSF) [43] to train the tabular dataset on the progression endpoint(s). RSF is an extension of Breiman's random forest algorithm to the time-to-event setting. In Appendix C, we include additional results using a penalized Cox proportional haz-

ards model with a ℓ_2 penalty (analogous to ridge regression) and a ℓ_1 penalty (analogous to lasso regression). All analyses in this section were conducted using R (version 4.2.1). RSF models were implemented using the package `randomforestSRC` [44] and the Cox models were implemented using the package `glmnet` [45].

In order to evaluate the quality of predictive information yielded via LOCO-MP, we trained the survival models on several distinct feature groupings of conventional and DBM features, outlined below:

1. **Conventional-only:** this model consists solely of the conventional features introduced in Section 2.1.
2. **All DBM:** this model consists of *all* DBM features (median and standard deviation of each region), after removing features with low variance (< 0.01). No conventional features are included.
3. **Conventional + All DBM:** this model uses all features from the "Conventional-only" and "All DBM" feature sets.
4. **Top DBM:** this model consists only of the top selected DBM features from the LOCO-MP algorithm, ranked by their feature occlusion scores $\bar{\Delta}_j$.[3] No conventional features are included.
5. **Conventional + Top DBM:** this model uses all conventional features and includes the top selected DBM features from the LOCO-MP algorithm, ranked by their feature occlusion scores $\bar{\Delta}_j$.

The first three feature groupings assess the individual and combined performance of the conventional and DBM features. The last two feature groupings are designed to evaluate the predictive potential of the DBM features selected by LOCO-MP.

The models were trained on each grouping using six repeats of 5-fold cross-validation. Model performance was measured by computing the Harrell's Concordance index (C-index). The C-index is analogous to the area under the ROC curve (AUROC) for binary classification tasks by comparing the observed time-to-event with the predicted risk and computing the proportion of patients where the two values are consistent (e.g. predicting higher risk for a patient with observed progression prior to another patient). A C-index of 1 denotes perfect model discrimination between patients, and a C-index of 0.5 indicates a fully random

---

[3] For illustration, we use the top six LOCO-MP features and discuss the motivation for this choice in the results section. We later demonstrate that the specific number of features used for modeling is less important than identifying a broad set of features for prediction.
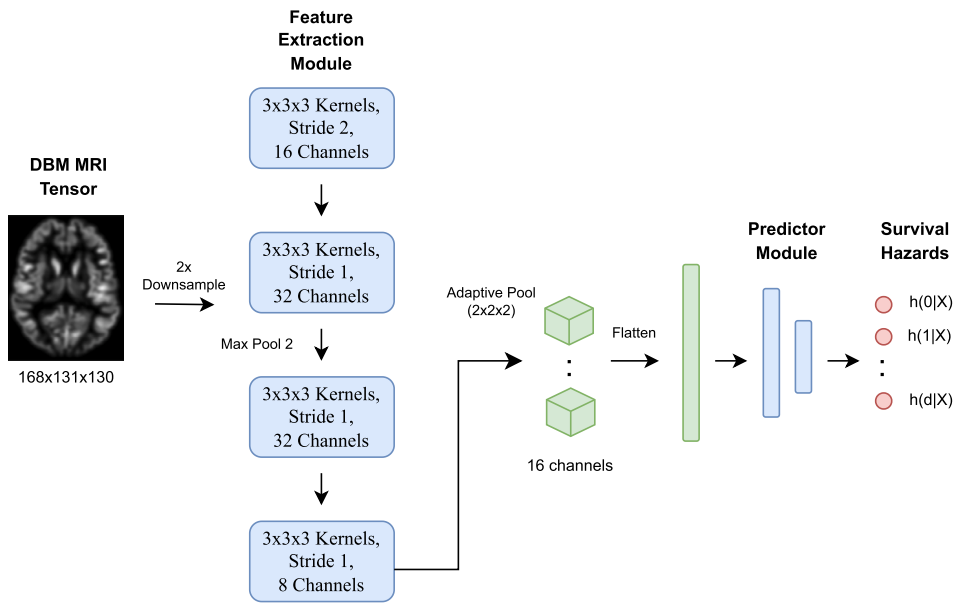
Fig. 3. Architecture diagram of the shallow full brain 3D CNN model. 3D convolutional layers extract local volumetric features from the MRI DBM voxel tensor. Summary features from 8 sections of the tensor are collected and passed to a feed-forward predictor module to output the discrete conditional hazards.

model. The primary outcome of our analysis is CCDP24, discussed in Section 2.1. In Section 3.2, we consider an alternative definition of progression based on the 25-foot walking test (S25FW).

2.4. MS progression with 3D convolutional neural networks

In contrast to the summary statistic featurization of the DBM data, 3D CNNs allow for flexible learning of localized features from the full brain or atlas region voxel-level tensors. Several applied works in MS have considered shallow 3D CNNs [22,20,46]. These relatively "parsimonious" architectures can be better suited to tasks with small sample size and without pretrained neural network weights to initialize the model with.

Our 3D CNN survival analysis pipeline, as applied to the full brain DBM tensor, is shown in Fig. 3. After the original full brain tensor is downsampled, it is passed through four 3D convolutional layers with kernel size 3, followed by a spatial adaptive max pool. The features resulting from the pool are flattened to a vector and passed through a two layer fully connected predictor module, which outputs discrete survival hazards. The network is trained using the same log-likelihood loss function as described in Equation (A.2). All neural networks are trained using `Pytorch 2.3.0` and `Python 3.10.14`.

We also train two additional CNN models from the literature, a deep convolutional attention model (CoAtNet), and a hierarchical transformer model designed for medical segmentation tasks (UNesT) [30,31]. A 3D implementation of CoAtNet is used [47] and it is pretrained using ImageNet [48]. For UNesT, the model is pretrained using T1-w MRIs. For both models, the final layer of their feature extractor modules are fed to an analogous predictor module as the in-house shallow architecture in Fig. 3.

Finally, we propose a novel convolutional module to process a user selected subset of the atlas regions. This region-level convolutional model, which we refer to as *Region CNN*, permits the user to use domain knowledge or previous feature selection work to select a sparse set of regions for the CNN to completely focus on, rather than processing the full brain. The architecture for *Region CNN* is depicted in Fig. 4. The user first specifies which regions are included in the model and separate CNN weights are then learned for each region. If the region is larger than $16 \times 16 \times 16$, it is cropped into regions of interest (ROIs) of this size, only keeping ROIs with sufficient ($> 30\%$) nonzero voxels.

Otherwise, the region tensor is left as a single ROI. The region ROIs are passed through an analogous set of four 3D CNN layers as described in Fig. 3. The only difference is that the stride of all CNN layers is 1. The resulting ROI feature vectors are collapsed to a single region-level feature vector using attention pooling [49]. Finally, the region-level feature vectors are concatenated to form the DBM feature vector for the patient. For model training, this vector is processed through a predictor module in the same fashion as Fig. 3.

3. Results

We begin by discussing our feature selection results from LOCO-MP and compare this against a traditional random survival forest model (Section 3.1). Sections 3.2 and 3.3 discuss the prediction modeling from traditional survival models and convolutional neural networks, respectively.

3.1. LOCO-MP feature selection

LOCO-MP identifies the most important DBM features for prediction in a stable manner. Each minipatch has dimension $n = N/5$ and $m = \sqrt{M}$, where $N = 350$ and $M = 56$. LOCO-MP was first applied to the entire dataset with $K = 10000$ minipatches. RSF and CCDP24 were used as the prediction model and progression outcome, respectively. The first four columns of Table 2 show the regions, summary statistics (i.e., the features), feature importance scores $\bar{\Delta}_j$, and rankings of the top six features identified from applying LOCO-MP to the full dataset. Fig. B.12 in the Appendix shows this for all 56 features. In the following paragraphs, we demonstrate the robustness of these feature importance results through various stability checks.

Stability of LOCO-MP In addition to the theoretical guarantees enjoyed by the LOCO-MP framework, we empirically validate its stability by testing its robustness against various data perturbations [50,51]. We first assess the robustness of the feature importance list by generating ten 80% subsamples of the data and applying LOCO-MP to each subsample, collecting the corresponding feature importance scores ($\bar{\Delta}_j$) and rankings. This is shown in the latter two columns of Table 2. Despite a low signal-to-noise ratio in the DBM data, the top six features are identical
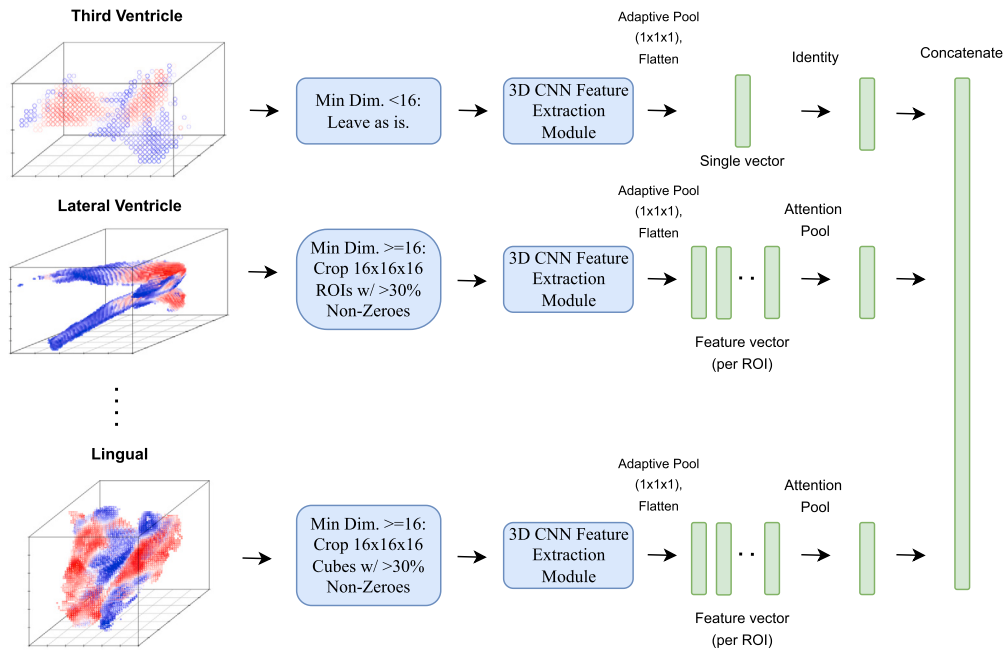
Fig. 4. Architecture diagram of the region-based 3D CNN feature extraction. For each atlas region, the DBM voxels are cropped into $16 \times 16 \times 16$ regions of interest (ROIs) comprising at least 30% nonzero values, if the region is larger than this crop size. The ROIs are passed through a 3D CNN feature extractor, following the specs of the feature extraction module in Fig. 3. The resulting tensors are spatially pooled and flattened to a feature vector for each ROI. The ROI feature vectors are pooled to a region-level feature vector using attention. Finally, the region-level feature vectors are concatenated to form the final DBM feature vector passed into the survival hazard predictor module from Fig. 3. A separate feature extractor is learned jointly for each selected atlas region.

Table 2
Top six LOCO-MP selected features based on full data and ten subsamples. The feature occlusion scores ($\bar{\Delta}_j$) and full data/median subsample ranks are also shown.

Region	Feature Summary Statistic	Full Data		Ten Subsamples	
		Rank	$\bar{\Delta}_j$	Median Rank	Median $\bar{\Delta}_j$
3rd Ventricle	Std. Dev	1	0.00104	2	0.00090
Lateral Ventricle	Median	2	0.00101	1	0.00095
Precuneus	Median	3	0.00088	3	0.00078
Cerebellum White Matter	Median	4	0.00056	4	0.00064
Lingual	Median	5	0.00051	5	0.00061
Parahippocampal	Median	6	0.00051	6	0.00042

Table 3
Summary of Jaccard index J for increasing numbers of top features.

Number of top features	2	3	4	5	6	7	8	9	10	11
Average J	0.93	0.85	0.84	1.00	0.94	0.80	0.89	0.77	0.72	0.68
Median J	1.00	1.00	1.00	1.00	1.00	0.75	0.89	0.80	0.67	0.69

between the full data and subsampled data models. The first two features are reversed in rank, but the difference in magnitude is extremely small (less than 0.00003).

We measure the consistency of the feature rankings by computing the *Jaccard similarity index J* between the ranks derived from the full dataset and those from the subsamples. This measures the proportion of overlap between two sets, ranging from 0 (no overlap) to 1 (complete overlap). By calculating *J* for varying numbers of top-ranked features, we aimed to quantify how reliably the top features appeared in the subsampled rankings. Table 3 presents the results for the top 11 features from the full data model. Across the ten subsamples, the average and median values of *J* decrease after six features before starting to plateau after ten. Therefore, we choose six as a heuristic number of DBM features to include in our model, though we demonstrate in Section 3.2 that our results remain relatively robust to varying numbers of top selected features.

Permutation test We tested the individual importance of the top DBM features by permuting their values across all patients and evaluating the ranks. For each feature, we permuted its values 25 times across patients and re-applied LOCO-MP for each permutation to evaluate the ranking of the permuted feature. The permutation disrupts the connections between individual DBM features and the outcome and tests whether higher ranks are attributable to specific feature identities. If the ranks of the permuted features are close to those from the unpermuted data, it would imply that the observed importance of the feature is not driven by the uniqueness of the feature (and corresponding brain region) itself. Instead, the perceived feature importance may be due to random and/or other global patterns in the data. The permuted ranks for each feature form a "pseudo-null distribution" under the null hypothesis that the feature is uninformative (shown in Fig. 5). This allows us to assess how often permuted ranks are as low or lower than the original rank. The results are shown in Fig. 5 – note that "Med" stands for the voxel
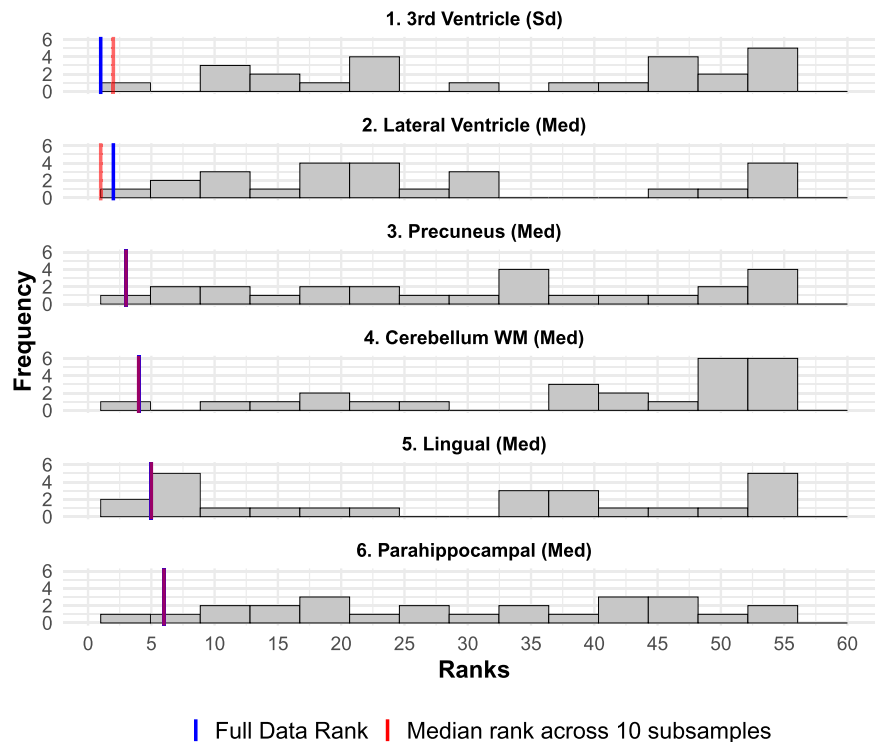
Fig. 5. LOCO-MP permutation test. We plot the distribution of the top six LOCO-MP features after 1) permuting their identities and 2) re-applying LOCO-MP to each permutation. The gray histograms show the rank distribution of each of the top LOCO-MP features after it is permuted. "Med" stands for voxel median and "Sd" stands for voxel standard deviation. The blue and red vertical lines represent the full data and median subsampled ranks (both unpermuted), respectively. These lines appear as a single purple line when the permuted and unpermuted ranks are identical.

median feature and "Sd" stands for the voxel standard deviation feature. For the top features identified by LOCO-MP, the ranks from the unpermuted models consistently fall below most, if not all, of the permuted ranks, yielding a low "empirical p-value." This suggests that the feature identities are indeed meaningful and unlikely due to chance, highlighting their potential relevance as medically important regions for MS progression, discussed in the last paragraph of this section.

LOCO-MP vs random forest-based scoring To assess the effectiveness of LOCO-MP, we benchmark it against the default feature importance scores produced by the random survival forest model which we refer to as "RF-Imp." This comparison illustrates the differences between our LOCO-MP approach and the results one might obtain by directly applying standard survival modeling software without first considering the underlying assumptions. While such tools are easily accessible, they may not be tailored for the challenges of high-dimensional feature selection.

The feature importance scores in RF-Imp are generated as the average difference in out-of-bag prediction error across decision trees when the corresponding feature is permuted. In principle, the bagging and feature subsets for each decision tree share some similarities with mini-patch ensembling. Incorporating shuffled feature values onto already-trained decision trees can create large erroneous changes to predictions that would otherwise have fine performance if retrained without the feature. This may introduce unnecessary instability to RF-Imp when compared with LOCO-MP. Following the same procedure as LOCO-MP, we first compute RF-Imp feature importance scores from the full dataset and then on ten subsamples. Fig. 6 shows the top six RF-Imp features based on the full data ranking. The blue boxplot shows the rank distribution across the ten subsamples, and the orange boxplot above it shows the corresponding LOCO-MP subsampled rank distribution for comparison.

Although there is alignment in the set of top regions captured between RF-Imp and LOCO-MP, the RF-Imp results alone demonstrate

larger rank instability, especially beyond the third and lateral ventricles. In particular, the 5th and 6th ranked features differ between LOCO-MP and RF-Imp, but the top four are the same. It is clear that LOCO-MP has considerably lower variability in its rank distributions over the ten subsamples, establishing it as a more stable choice for feature selection. In contrast, the RF-Imp ranks have greater spread across subsamples, increasing the possibility of identifying spuriously meaningful features. As an additional stability check, we replace random forest with gradient boosting machines (GBM) as the underlying survival algorithm in LOCO-MP and find that the subsample ranks of the top regions are concordant with RF, excluding the parahippocampal region (Table B.5). The greater stability of LOCO-MP allows us to be more confident that the selected features are genuinely meaningful and not due to noise or other artifacts of the data. In future work, an ensembling of LOCO-MP with multiple survival models may yield an even more stable feature set.

Similar top features are identified when LOCO-MP is combined with survival gradient boosting machines (GBM) instead of random forest, highlighting that LOCO-MP can be relatively algorithm agnostic.

Clinical relevance of identified regions In spite of the modestly sized LOCO-MP feature importance scores arising from the small cohort MRI data, the features and their ranks are stable. These features also align with brain regions that are clinically relevant in MS. The two identified ventricular regions, including the third and lateral ventricles, are often indicated in MS progression studies [52–54]. The consistency of ventricle-related findings across different studies emphasizes the robustness of these regions. Notably, these regions were also identified as significant in a recent causal study using the treatment arm of the same trial [6]. While this study focused on treatment effects, our use of the control arm to predict progression confirms the continued importance of the ventricular regions even in the absence of therapeutic intervention.
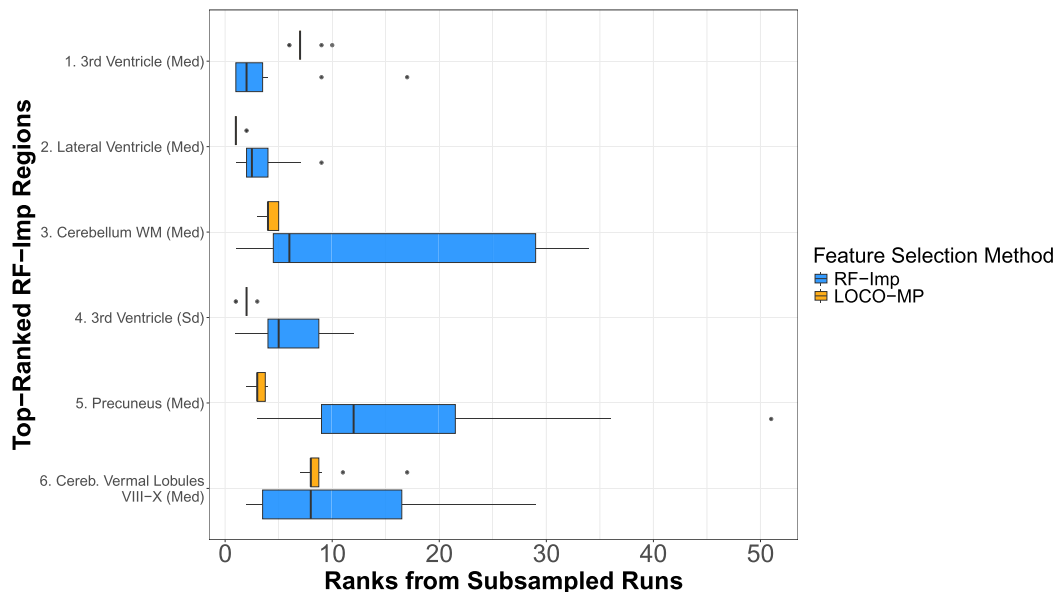
Fig. 6. Rank distribution for RF-Imp feature importance (blue) vs LOCO-MP (orange). The *y*-axis shows the top six RF-Imp features in terms of their feature importance score from an RSF model applied to the full data. The *x*-axis plots the ranks across different data subsamples. "Med" stands for voxel median and "Sd" stands for voxel standard deviation. The blue boxplots show their subsampled ranks on the *x*-axis. The orange boxplot above to each blue boxplot shows the rank distribution of that feature from LOCO-MP.

The cerebellum is also frequently indicated in MS studies [55,56], which may play a role in impairment of postural control and balance in MS patients [57]. The precuneus and lingual gyrus are found abnormally activated in the brains of MS patients when performing cognitive exercises, and this activation was later associated with increased mental fatigue and slower speed in task completion [58]. Finally, the parahippocampal gyrus (ranked 6th) has also been linked to varying activation patterns in MS patients compared to controls [59], and is also associated with much greater lesion activity compared to controls [60].

While the identified regions possess some degree of clinical relevance in previous MS studies, we emphasize that our study is exploratory in this area and does not intend to draw any clinical conclusions. Pathological development of MS can be diffusive across brain anatomy and functions [1,61]. As a result, many brain regions can atrophy or be implicated by the disease. In order to inform future clinical decision-making, further validation on larger and more heterogeneous patient populations is required. However, our study provides a practical first step towards future scientific studies on certain brain regions.

3.2. Prediction improvements from selected regions under classical survival models

We show that the predictive models comprising LOCO-MP identified features have stronger discriminative power than models using all DBM features. We first present results from CCDP24, a canonical endpoint for measuring MS progression. We then discuss how the LOCO-MP selected features generalize well to alternative endpoints and compensate for the loss of key conventional predictors.

3.2.1. CCDP24 outcome

Our RSF model results for the CCDP24 outcome are presented in Fig. 7. Each boxplot shows the distribution of test-set C-indices for a specific feature grouping (discussed in Section 2.3) across all cross-validation folds and repeats. The median test-set C-index is annotated within each boxplot. Unsurprisingly, the "All DBM" model has the lowest C-index across the feature groups, likely due to overfitting from the high dimensionality of using all 56 DBM features. A similar conclusion can be drawn from the "Conventional + All DBM" model which includes
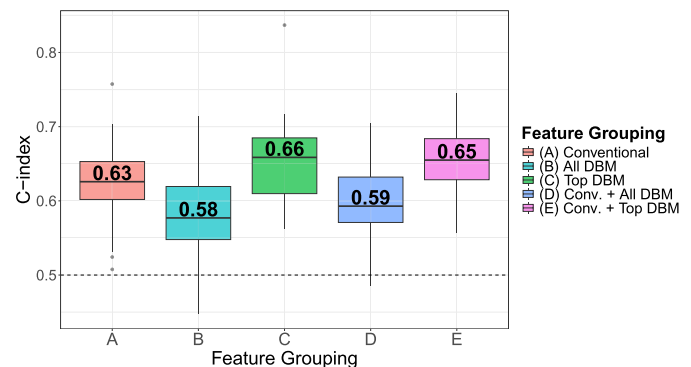

Fig. 7. CCDP24 RSF model results. Boxplots show the distribution of random survival forest test-set C-indices across all cross-validation folds and repeats for outcome CCDP24. The *x*-axis consists of the feature groupings and the *y*-axis shows the test set C-indices. "Top DBM" refers to the top six DBM features. The annotated C-index represents the median value for a specific feature grouping.

even more features, showing how using an extensive set of MRI features without clinical context can lead to suboptimal predictive performance. These two models were slightly outperformed by the model using conventional features only, suggesting that the conventional features alone capture some degree of predictive value in CCDP24. In contrast, the two models that include the top features selected from DBM demonstrate the best prediction performance. Both C-indices are centered around 0.67, suggesting moderate predictive ability from the top DBM features. Adjusting for clinical information in the "Conventional + All/Top DBM" models did not appear to affect performance.

For simplicity, the "Top DBM" models (Top DBM, Conventional + Top DBM) shown in Fig. 7 are fit with the top six LOCO-MP features. However, the results from these two feature groupings remained stable with respect to varying numbers of LOCO-MP-selected features. Fig. 8 plots the distribution of test-set C-indices from these two feature groupings across the top-ranked DBM features, ranging from 1 to 8. Performance remains stable across different numbers of features, suggesting that the model is robust to how many top features are used – the precise
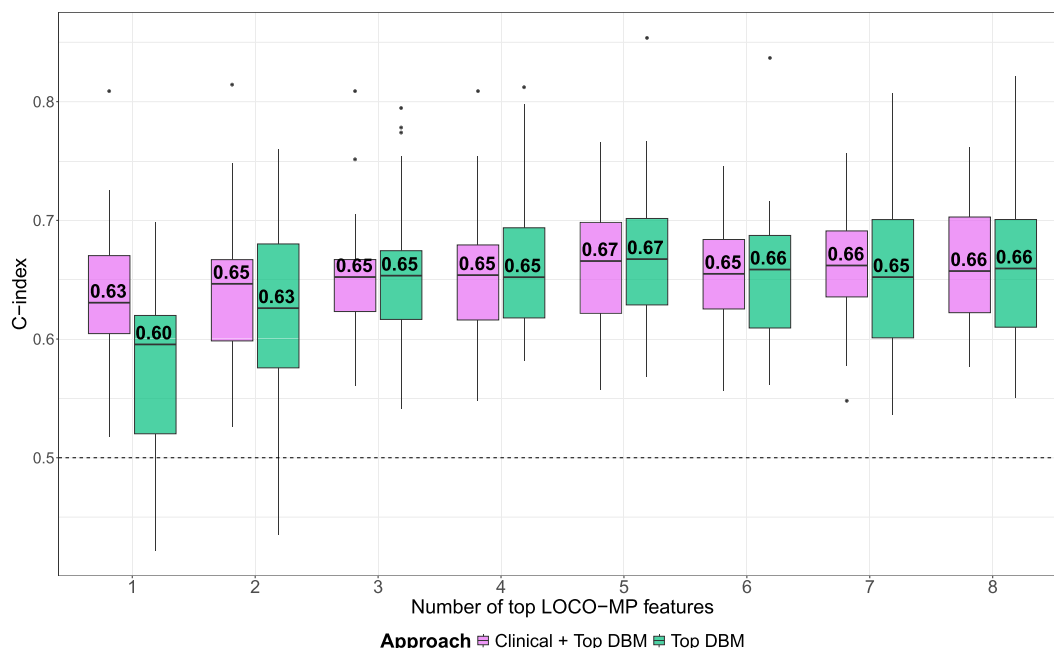
Fig. 8. CCDP24 model stability. Boxplots show the distribution of CCDP24 test-set C-indices for the Conventional + Top DBM and Top DBM-only RSF models, plotted across varying numbers of top LOCO-MP features. The *x*-axis represents the number of top LOCO-MP features selected and the *y*-axis is the C-index.

rankings between features and the specific number used for modeling is less important than identifying a broad set of features that can be investigated for clinical relevance. Although these two feature groupings are more investigative, researchers can be confident that the results are not overly sensitive to the specific subset of features. To assess generalizability of the Top DBM features, Section 3.2.2 applies this feature set to an alternative progression outcome, and Appendix D applies this feature set to a separate control cohort.

Several conclusions can be drawn from our results. First, many of the DBM features contain overlapping information that ultimately results in model overfitting, as evidenced by the two "All DBM" model boxplots. This underscores the value of LOCO-MP, which can sift through highly correlated groups of features to identify the most prognostic ones and improve prediction performance. Moreover, the predictive strength of the DBM features and the clinical relevance of the corresponding brain regions reinforces the validity of these findings, demonstrating that LOCO-MP identifies DBM features with genuine predictive value.

3.2.2. 25-foot walking test outcome

Despite variations in performance from across different feature sets, the CCDP24 C-indices are modest for even the strongest feature groups. This may arise from various factors, including inherent limitations in the outcome definition which impacts the model's ability to identify meaningful signal – CCDP24 is characterized by marked changes in multiple events that are sustained for 24 weeks [62,63]. Although combining multiple events reduces the censoring rate, it comes at the cost of potentially introducing subjective measurements into the analysis: EDSS scoring can vary across studies, and the definition of EDSS-based progression highly depends on the study at hand [64]. Therefore, including information from noisy outcome measurements could potentially worsen performance and we hypothesize this may be the case with CCDP24. To combat this, recent literature has pointed to the timed 25-foot walking test as an alternative measure of disability progression that may be more sensitive in detecting genuine disease worsening [64]. This test measures the time for a patient to walk 25 feet without any assistive devices. Progression is defined as a 20% increase in 25-foot walking time that is sustained for 24 weeks (S25FW). Previous studies have demonstrated significantly worse S25FW times in MS patients [65] and S25FW
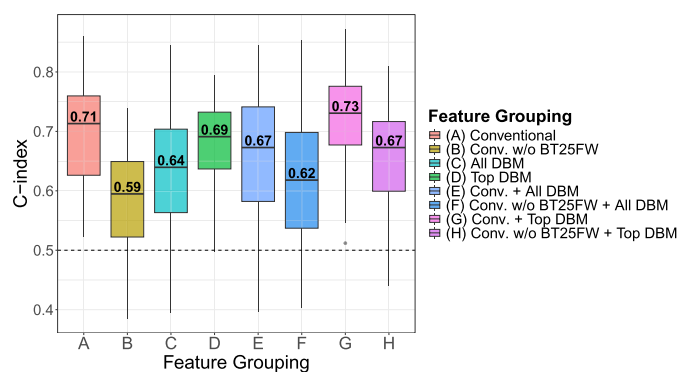

Fig. 9. S25FW RSF model results. Boxplots show the distribution of random survival forest test-set C-indices across all cross-validation folds and repeats for outcome S25FW. The *x*-axis consists of the feature groupings and the *y*-axis shows the test set C-indices. "Top DBM" refers to the top six DBM features. The annotated C-index represents the median value for a specific feature grouping. We also include results after omitting the highly prognostic conventional variable BT25FW.

has previously been shown to be strongly correlated with other forms of MS progression [66].

This outcome has a higher censoring rate compared to CCDP24 (88% compared to 77%), but resulted in higher C-indices across all feature groupings, shown in Fig. 9. Similar to CCDP24, a model using all DBM features leads to overfitting. However, the conventional feature model demonstrates a 12.7% increase in median C-index compared to CCDP24 (0.63 to 0.71), and the "Conventional + Top DBM" and "Top DBM-only" models also see increases in median C-index. These increases suggest that DBM information is more effective in predicting S25FW-based progression. This also supports previously stated claims about the improved sensitivity of S25FW, illustrating the perils of using a noisy outcome measurement like CCDP24. The C-index for S25FW also remained robust to different numbers of top DBM features, as shown in Fig. 10.

The strong performance in the conventional model for S25FW is largely attributed to the baseline 25-foot walking test feature (BT25W,
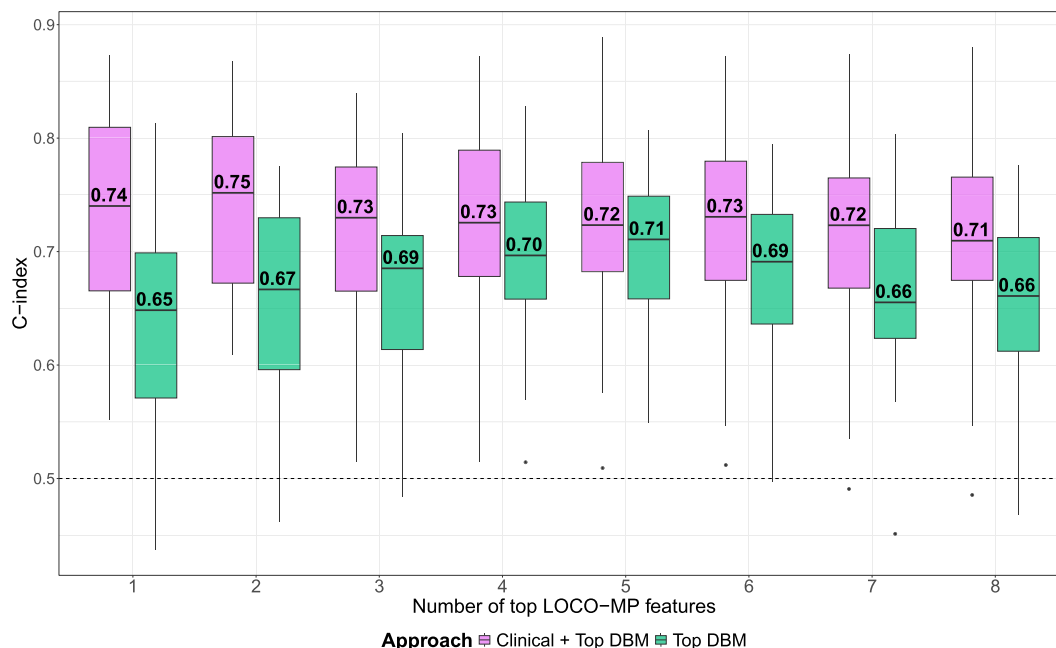
Fig. 10. S25FW model stability. Boxplots show the distribution of S25FW test-set C-indices for the Conventional + Top DBM and Top DBM-only RSF models, plotted across varying numbers of top LOCO-MP features. The *x*-axis represents the number of top LOCO-MP features selected and the *y*-axis is the C-index.

see Section 2.1) which is highly predictive of S25FW progression. The C-indices decrease after omitting BT25W as a conventional covariate in our models – the "Conventional-only" and "Conventional + All DBM" models achieve a median C-index of 0.59 and 0.62, respectively, as shown in Fig. 9. These performances are comparable to their CCDP24 counterparts. While omitting BT25FW results in worse performance across all feature groupings with conventional features, the "Conventional + Top DBM" model performs fairly robustly even after omitting BT25FW, achieving a median C-index of 0.67. This demonstrates that the top DBM features may be able to compensate for the loss of key conventional predictors and shows how LOCO-MP can identify robust and generalizable features.

We also note that the two "Top DBM" models in the S25FW analysis use LOCO-MP features identified from the CCDP24 outcome – the algorithm encountered challenges when applied to S25FW due to its high censoring rate, which limited the number of informative minipatches (many minipatches consisted entirely of censored patients). Despite this, the increased performance in the two "Top DBM" S25FW models demonstrates the reliability of LOCO-MP in identifying a stable feature set that captures fundamental aspects of disease progression that are robust across different outcome definitions. Appendix D further demonstrates that these identified DBM improve S25FW prediction in a separate validation cohort. These strong generalizations are further evidence that the top features may be genuine markers of disease worsening, enhancing its potential for broader clinical applications and underscoring the value of DBM as a regional identification framework.

While this new outcome allowed for more precise patient risk discrimination, the variance of the C-index estimates was higher than that of CCDP24. This is due to the higher censoring rate in S25FW which, by reducing the number of informative events, lowers the effective sample size. This generally leads to increased variance in the estimates because there is less observed data available to accurately capture the underlying risk. Even though S25FW provides clearer risk stratification, the higher variation in the C-index reflects the statistical challenges introduced by a lower effective sample size. In contrast, the CCDP24 outcome has lower variance but lower C-indices as well.
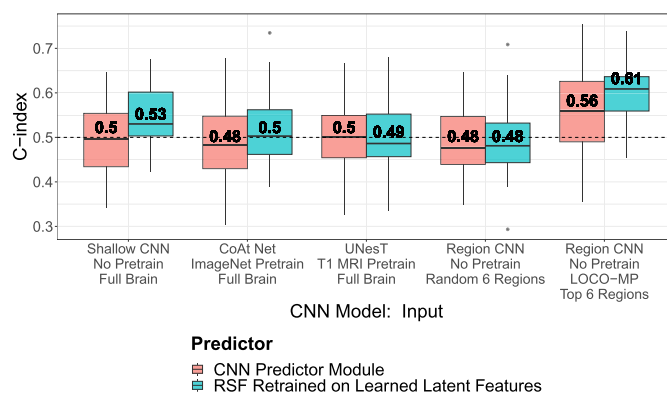

Fig. 11. CNN model performance for CCDP24. Boxplots show the distribution of test C-indices for each of the tested CNN models. Results are presented both for the trained CNN predictor module (CNN Predictor Module), as well as an RSF trained on the extracted feature vector of each trained CNN (RF Retrained on Extracted Features). The *x*-axis labels first give the CNN model architecture used, followed by the inputs used. Specifically, full brain models are compared with *Region CNN*, using either the top 6 LOCO-MP regions or random 6 regions (excluding 6 LOCO-MP regions) with each fold.

3.3. Prediction improvements from selected regions under convolutional neural networks

We compare full brain CNNs against *Region CNN* using only the top 6 identified LOCO-MP regions to assess the extent to which an unsupervised deep learning feature extraction improves prediction performance. Fig. 11 presents CNN model test C-indices over 30 80%/20% data splits for predicting CCDP24. The evaluated models include the three baseline full brain models (Shallow, CoAtNet UNesT), *Region CNN* trained on the top 6 LOCO-MP regions, and *Region CNN* trained on 6 randomly selected regions (excluding the top 6 LOCO-MP regions). Each of these models are described in Section 2.4. Red boxplots show the C-index distribution from the predictions of the CNN predictor module. The learned latent features from each CNN (i.e., the concatenated green bar

in Fig. 3) were also extracted and retrained using a conventional RSF survival model (without conventional features), enabling the continuous survival times to be used as the outcome. Teal boxplots show the C-index distribution of the RSF models trained on these extracted CNN feature vectors.

These results underscore the importance of subsetting the DBM input to the top 6 LOCO-MP regions in order to extract signal from a CNN. Each of the full brain CNN models struggle to achieve nonrandom performance (C-index ≥ 0.5). The shallow full brain CNN features perform modestly better than the complex CoAtNet and UNesT architectures. This suggests that, for this data and cohort size, neither an attention mechanism nor a very deep architecture are able to adequately predict CCDP24. Additionally, CoAtNet was pretrained using ImageNet [48] and UNesT was pretrained using raw T1-w MRI tensors, indicating that the DBM data may reflect critically different patterns than the raw MRI. *Region CNN* fares quite poorly when trained on 6 random non-LOCO-MP regions. However, it extracts strong predictive signal when LOCO-MP features are used. The LOCO-MP features were more helpful in the RSF model, achieving a median C-index of 0.61. *Region CNN* with the top 6 LOCO-MP regions performs comparably on the S25FW outcome (median predictor module C-Index: 0.57; median RSF retrained C-Index: 0.61). Ultimately, even the best CNN fares moderately worse than the corresponding classical survival model in Section 3.2, highlighting the challenge of learning complex models for the DBM data on this cohort size. With more data, *Region CNN* (and CNNs generally) may be able to extract more expressive features from the T1-w MRI data.

4. Discussion

This study develops a data analysis pipeline that balances stability, predictive performance, and domain-level interpretability to improve MS prediction models with DBM-based features. We demonstrate how LOCO-MP identifies clinically relevant features in a stable manner, outperforming other feature importance approaches. Our sample size of 350 is modest for complex prediction tasks such as MS progression prediction from neuroimaging data. While sample size can influence the stability and generalizability of model outputs, we mitigated these concerns by conducting extensive stability checks discussed in Section 3.1 and varying the number of top DBM features used for modeling (Figs. 8 and 10). The selected LOCO-MP features generalized well to a separate validation cohort Appendix D. We note that our analysis pipeline is scalable and well-suited to larger datasets as they become available.

Regarding prediction performance, we observe that the test set C-indices for the strongest feature groupings primarily range between 0.65-0.75. We note that our prediction task focuses on 24-week progression, a relatively short time frame for a chronic illness such as MS. Predicting over such brief periods is inherently challenging, as disease progression tends to be subtler and more variable, limiting the ability of models to capture robust patterns. This limitation likely prevents the observed C-index from being higher. In fact, several other RRMS patient studies that predict short-term progression with similar sample sizes report comparable C-indices or AUROC values[4] (between 0.48 and 0.73) when using a similarly-defined outcome with different types of brain data [18,67–69]. Another study reported slightly higher C-indices (0.76) when integrating genotypic factors with clinical variables [70] – the use of non-MRI based features, such as those based on genetic or gut microbiome information, has proven to yield strong predictions and is a potential future direction for MS prediction [71]. In addition, Pellegrini et al. (2020) observed that C-indices did not exceed 0.65 with conventional survival models, despite using a larger cohort [18].

The endpoints themselves also have limitations. As mentioned in Section 3.2.2, CCDP24 is derived by combining the scores of EDSS, 9HPT, and T25FW. While each of the three clinical assessments has its own limitations [72–74], EDSS is more subjective and less sensitive to disease-related changes [72]. S25FW is a quantitative measure of lower limb movement function. Because it only captures one specific function, its two-year progression rate experiences more censoring. Progression on this metric regardless has a high correspondence with overall RRMS progression [73]. It remains an area of future work to find a clinical progression measure that optimally balances sensitivity and censoring rates.

We additionally explored the performance of 3D CNN models using the full brain versus the LOCO-MP top regions. Though the 3D CNN models substantially improved when constraining the DBM input to the LOCO-MP top DBM regions, their performance still lagged behind conventional modeling using summary statistics of these regions. State of the art 3D CNN architectures did not change this story; the small sample size could be the main limiting factor for deep learning performance. In the future, we are interested in a deeper exploration of small sample size strategies for CNNs, such as DBM pretraining with self-supervised learning (SSL) on DBM-processed T1 MRI data, model ensembling, domain-tailored attention, and few-shot learning techniques.

Ultimately, our results suggest that DBM uniquely captures subtle volumetric abnormalities across the brain and provides orthogonal information to the conventional features, contributing to MS progression in a way that is stable and clinically meaningful across multiple definitions of progression. Future work could integrate DBM with other MRI modalities used for lesion localization. Additionally, our study was restricted solely to patients with relapsing MS. Future work could utilize DBM to differentiate between different subtypes of MS or predict the conversion of relapsing MS to secondary progressive MS [75] when those subtypes are available. Our analysis pipeline is an initial starting point for more advanced data-driven studies that may provide better scientific and clinical insights about how MS impacts the brain.

CRediT authorship contribution statement

Andy A. Shen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Aidan McLoughlin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Zoe Vernon:** Visualization, Methodology, Formal analysis. **Jonathan Lin:** Software. **Richard A.D. Carano:** Supervision, Project administration. **Peter J. Bickel:** Writing – review & editing, Supervision, Project administration. **Zhuang Song:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Investigation, Data curation, Conceptualization. **Haiyan Huang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

Richard A.D. Carano and Zhuang Song are employees of Genentech/Roche and are stockholders of Roche.

The authors have no conflicts of interest to report.

Acknowledgements

---

[4] The C-index is analogous to the AUROC in binary prediction tasks with survival data.

Appendix A. Details of LOCO-MP for survival analysis

Following the notation of Gan et al. (2023) [28], we denote the $N \times M$ data matrix as X and the outcome as Y. Let (X_i, Y_i) denote the feature-outcome pair for individual i, where X_i is a M-dimensional vector and $Y_i = (T_i, C_i)$. For every minipatch $k = 1 \ldots K$, we sample n observations I_k and m features F_k from X. Denote this minipatch as X_{I_k, F_k}. A prediction model $\hat{\mu}_k$ is then trained on X_{I_k, F_k} and predictions are generated for the remaining observations X_{-I_k, F_k}, which includes features F_k and observations outside I_k. Repeating this over all K minipatches, we compute the individual feature occlusion score for feature j and observation i:

$$\hat{\Delta}_{ij} = \text{Error}(Y_i, \hat{\mu}_{-i}^{-j}(X_i)) - \text{Error}\left(Y_i, \hat{\mu}_{-i}(X_i)\right),$$

where

$$\hat{\mu}_{-i}(X_i) = \frac{1}{\sum_{k=1}^{K} \mathbb{1}(i \notin I_k)} \sum_{k=1}^{K} \mathbb{1}(i \notin I_k) \hat{\mu}_k(X_i), \text{ and}$$

$$\hat{\mu}_{-i}^{-j}(X_i) = \frac{1}{\sum_{k=1}^{K} \mathbb{1}(i \notin I_k)\mathbb{1}(j \notin F_k)} \sum_{k=1}^{K} \mathbb{1}(i \notin I_k)\mathbb{1}(j \notin F_k) \hat{\mu}_k(X_i),$$

and Error() refers to any observation-level prediction error function, in our case the discrete hazard loss discussed in the following paragraph. The individual $\hat{\Delta}_{ij}$ values are then averaged, giving us the feature occlusion score for feature j:

$$\bar{\Delta}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}_{ij}. \tag{A.1}$$

$\hat{\Delta}_{ij}$ is the difference in error for observation i when feature j is omitted. Concretely, this can be thought of as the change in prediction capability for patient i when a specific DBM feature is omitted. Larger, positive values of $\hat{\Delta}_{ij}$ indicate greater feature importance since the model performs worse (has larger error) when feature j is excluded.

The prediction error for survival analysis models is commonly measured as the reciprocal of the overall concordance index (C-index) of risk scores between pairs of observations. However, this metric is not an aggregation of observation-level errors, preventing its direct usage in the LOCO-MP framework. To address this, we utilize discrete hazard loss functions used in deep learning survival analysis models [41,42,76]. Specifically, we evenly divide the survival time scale of the study into d intervals: $[t_0, t_1), \ldots, [t_{d-1}, t_d)$, where t_d is the end of the study period. The patient's event time is now denoted $T_i = q$ if and only if $T_i \in [t_q, t_{q+1})$.

Given this setup, we may define a patient's conditional hazard probability as:

$$h(q \mid X_i) = \mathbb{P}(T_i = q \mid T_i \geq q, X_i).$$

This patient's survival probability follows as the probability of surviving until the end of the current interval or longer:

$$S(q \mid X_i) = \mathbb{P}(T_i > q \mid X_i) = \prod_{s=1}^{q} (1 - h(s \mid X_i)).$$

From these definitions, we can compute an observation-level likelihood. For uncensored patients, this is the product of the conditional hazard at their event time and the survival function of the prior time period:

$$\ell_U(T_i = q) = h(q \mid X_i) \, S(q - 1 \mid X_i).$$

For patients that are right-censored at time period q, their likelihood is simply the survival function at time period q:

$$\ell_C(T_i > q) = S(q \mid X_i).$$

Altogether, given the censoring status C_i for each patient, the observation level model error for LOCO-MP at time $T_i = q$ is defined as the negative log-likelihood of the discretized observation:

$$\text{Error}((T_i, C_i), \hat{\mu}_{-i}^{-j}) = \begin{cases} -\log(h(q \mid X_i)) - \log(S(q - 1 \mid X_i)); & C_i = 0 \\ -\log(S(q \mid X_i)); & C_i = 1 \end{cases} \tag{A.2}$$

Here, $\hat{\mu}_{-i}^{-j}$ encapsulates the collection of conditional hazards that are computed with the underlying minipatch model. Assuming that the patient conditional hazard probabilities are not exactly 0 or 1 over the study period, this loss framework satisfies the necessary assumptions for the theoretical guarantees of LOCO-MP [28].

LOCO-MP is advantageous over other feature selection techniques due to its ability to provide asymptotic inferential guarantees [28]. Moreover, LOCO-MP can also account for dependencies across features, a common issue in high-dimensional settings like ours – Gan et al. (2023) [28] discuss that, by generating randomly subsampled features across minipatches, LOCO-MP ensures that the predictive value of each feature is not being diminished by other strongly correlated features, since groups of strongly correlated features will not always appear in the same minipatch. The full LOCO-MP procedure and its theoretical guarantees are discussed in Gan et al. (2023) [28].

Appendix B. Additional LOCO-MP and feature selection results

B.1. Full LOCO-MP results

Fig. B.12 shows boxplots for all 56 DBM features, similar to Fig. 6 in the main text. Blue boxplots show the rank distribution from RF-Imp while orange boxplots show the rank distribution from LOCO-MP.

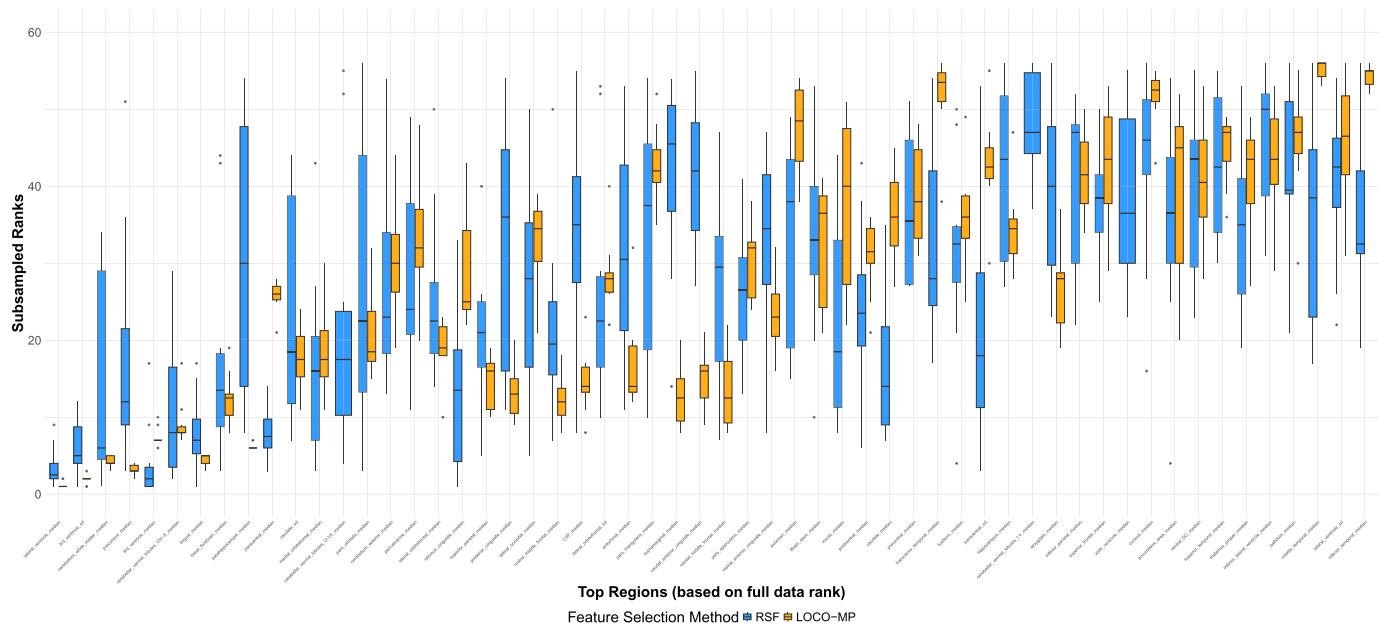Table B.4 shows the LOCO-MP feature importance scores for all 56 DBM features.

Fig. B.12. Rank distribution for RSF feature importance (blue) vs LOCO-MP (orange) for all DBM features. The *x*-axis shows the features (from highest to lowest rank) in terms of their feature importance score from an RSF model applied to the full data, and the blue boxplots show their subsampled ranks on the *y*-axis. The orange boxplot next to each blue boxplot shows the rank distribution of that feature from LOCO-MP.

Table B.4
Feature importance scores and rankings from full data and subsamples for all 56 high-variance VoxelDBM regions.

Region	Full Data		Subsamples	
	Rank	$\bar{\Delta}_j$	Median Rank	Median $\bar{\Delta}_j$
3rd Ventricle SD	1	0.00104	2	0.00090
Lateral Ventricle Median	2	0.00101	1	0.00095
Precuneus Median	3	0.00088	3	0.00078
Cerebellum White Matter Median	4	0.00056	4	0.00064
Lingual Median	5	0.00051	5	0.00061
Parahippocampal Median	6	0.00051	6	0.00042
Cerebellar Vermal Lobules VIII-X Median	7	0.00029	8	0.00027
3rd Ventricle Median	8	0.00028	7	0.00037
Basal Forebrain Median	9	0.00026	12.5	0.00016
Caudal Anterior Cingulate Median	10	0.00023	16	0.00014
Caudal Middle Frontal Median	11	0.00017	12.5	0.00019
Supramarginal Median	12	0.00016	12.5	0.00017
Rostral Middle Frontal Median	13	0.00016	12	0.00019
CSF Median	14	0.00015	14	0.00015
Lateral Orbitofrontal Median	15	0.00014	19	0.00008
Posterior Cingulate Median	16	0.00014	13	0.00017
Pars Orbitalis Median	17	0.00012	18.5	0.00008
Caudate SD	18	0.00010	17.5	0.00009
Medial Orbitofrontal Median	19	0.00005	17.5	0.00008
Paracentral Median	20	0.00004	26	-0.00002
Pericalcarine Median	21	0.00003	32	-0.00010
Superior Parietal Median	22	0.00002	16	0.00014
Cerebellum Exterior Median	23	-0.00000	30	-0.00007
Entorhinal Median	24	-0.00002	14	0.00014
Amygdala Median	25	-0.00002	28	-0.00003
Brain Stem Median	26	-0.00003	36.5	-0.00014

(continued on next page)

Table B.4 (*continued*)

Region	Full Data		Subsamples	
	Rank	$\bar{\Delta}_j$	Median Rank	Median $\bar{\Delta}_j$
Rostral Anterior Cingulate Median	27	-0.00003	23	0.00001
Lateral Occipital Median	28	-0.00005	34.5	-0.00011
Isthmus Cingulate Median	29	-0.00005	25	-0.00002
Pars Triangularis Median	30	-0.00006	42	-0.00021
Superior Frontal Median	31	-0.00007	43.5	-0.00021
Pars Opercularis Median	32	-0.00008	32	-0.00007
Lateral Orbitofrontal SD	33	-0.00010	28	-0.00005
Hippocampus Median	34	-0.00010	34.5	-0.00011
Cerebellar Vermal Lobules I-V Median	35	-0.00010	25.5	-0.00003
Caudate Median	36	-0.00012	36	-0.00012
Fusiform Median	37	-0.00012	36	-0.00015
Accumbens Area Median	38	-0.00013	45	-0.00022
Insula Median	39	-0.00019	40	-0.00020
Inferior Parietal Median	40	-0.00019	41.5	-0.00019
Postcentral Median	41	-0.00019	31.5	-0.00010
Pallidum Median	42	-0.00020	47	-0.00023
Ventral DC Median	43	-0.00020	40.5	-0.00018
Precentral Median	44	-0.00023	38	-0.00016
Transverse Temporal Median	45	-0.00023	53.5	-0.00035
Cerebellar Vermal Lobules VI-VII Median	46	-0.00024	49	-0.00026
Putamen Median	47	-0.00025	48.5	-0.00027
Lateral Ventricle SD	48	-0.00025	46.5	-0.00024
4th Ventricle Median	49	-0.00025	45.5	-0.00024
Inferior Lateral Ventricle Median	50	-0.00026	43.5	-0.00021
Superior Temporal Median	51	-0.00026	47	-0.00023
Paracentral SD	52	-0.00029	42.5	-0.00020
Cuneus Median	53	-0.00033	52.5	-0.00033
Thalamus Proper Median	54	-0.00035	43.5	-0.00022
Inferior Temporal Median	55	-0.00047	55	-0.00041
Middle Temporal Median	56	-0.00052	56	-0.00046

Table B.5
Ranks of median predictive importance across ten subsamples under GBM and GBM-LOCO. Results presented for top six LOCO-MP random survival forest features.

Region	Feature Summary Statistic	Median Subsample Rank	
		GBM	GBM-LOCO
3rd Ventricle	Std. Dev	6	2
Lateral Ventricle	Median	1	1
Precuneus	Median	18.5	7
Cerebellum White Matter	Median	5	5
Lingual	Median	9.5	13
Parahippocampal	Median	25	21.5

B.2. LOCO-MP under gradient boosting machines

Though we primarily study the performance of LOCO-MP under nonparametric random survival forests, the feature selection framework may utilize other machine learning survival algorithms, such as gradient boosting machines (GBM). We incorporate a survival GBM implementation [77] into LOCO-MP (GBM-LOCO), and capture its DBM feature importance ranks in OPERA I across ten 80% subsamples, analogously to the experiments in Section 3.1. We also capture the feature ranks of GBM alone, which are assigned based on the reduction in negative Cox partial log likelihood.

In Table B.5, for the top six LOCO-MP random forest features, we present the ranks of their median predictive importance across ten subsamples under GBM and GBM-LOCO. The importance of the third and lateral ventricles and the cerebellum white matter is retained for both GBM and GBM-LOCO. The GBM-LOCO ranks are more concordant with RF-LOCO than are the GBM ranks, particularly for the third ventricle and precuneus. The parahippocampal region is unimportant under both GBM and GBM-LOCO. This suggests that an ensembling approach of multiple machine methods under LOCO could help further validate brain regions that are stably identified.

Appendix C. Penalized Cox proportional hazards modeling

Figs. C.13 and C.14 show the distribution of test set C-indices across all folds for each feature group when a penalized Cox proportional hazards model (ridge penalty) is used to predict progression. Fig. C.13 uses the CCDP24 outcome and Fig. C.14 uses the 25-foot walking outcome. Figs. C.15 and C.16 show the analogous plots for the Cox proportional hazards model with a lasso penalty. (See Figs. C.17 and C.18.)
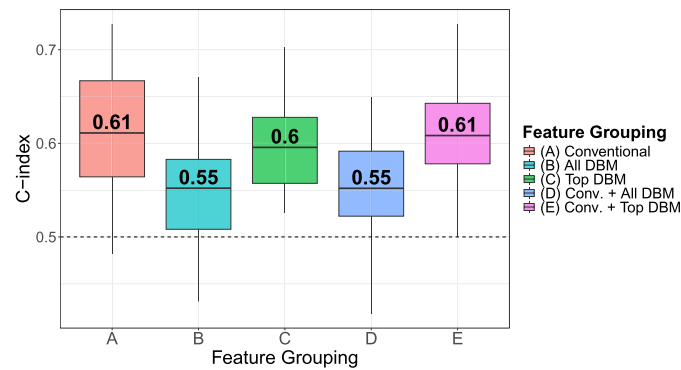
Fig. C.13. CCDP24 Cox model results (ridge). Boxplots show the distribution of penalized Cox proportional hazards model (ridge penalty) test-set C-indices across all cross-validation folds and repeats for outcome CCDP24. The *x*-axis consists of the five feature groupings and the *y*-axis shows the test set C-indices. "Top MRI" refers to the top six DBM features. The annotated C-index represents the median value for a specific feature grouping.
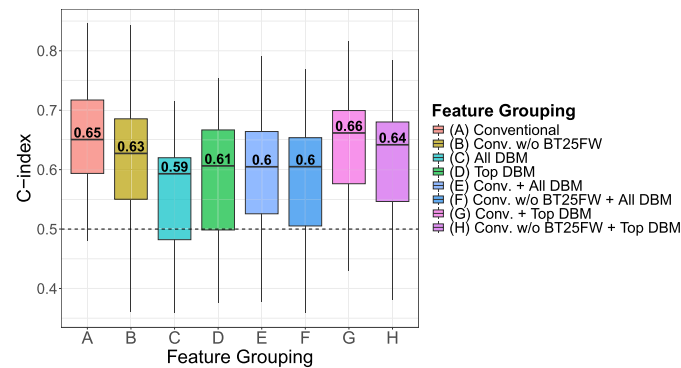

Fig. C.14. S25FW Cox model results (ridge). Boxplots show the distribution of penalized Cox proportional hazards model (ridge penalty) test-set C-indices across all cross-validation folds and repeats for outcome 25FW. The *x*-axis consists of the five feature groupings and the *y*-axis shows the test set C-indices. "Top MRI" refers to the top six DBM features. The annotated C-index represents the median value for a specific feature grouping.
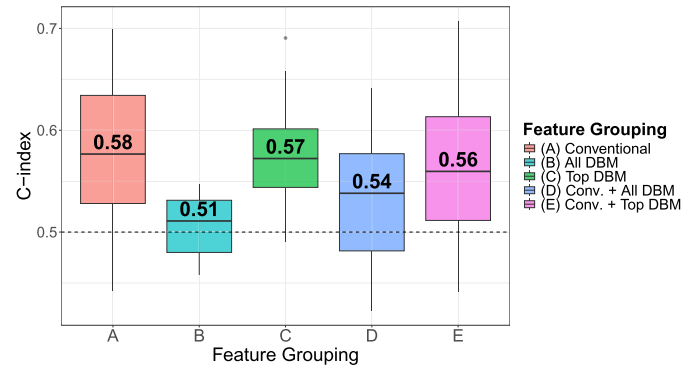

Fig. C.15. CCDP24 Cox model results (lasso). Boxplots show the distribution of penalized Cox proportional hazards model (lasso penalty) test-set C-indices across all cross-validation folds and repeats for outcome CCDP24. The *x*-axis consists of the five feature groupings and the *y*-axis shows the test set C-indices. "Top MRI" refers to the top six DBM features. The annotated C-index represents the median value for a specific feature grouping.

Figs. C.19 and C.20 show the stability of the test set C-indices across varying numbers of top LOCO-MP features for the penalized Cox proportional hazards model (ridge penalty). Fig. C.19 uses the CCDP24 outcome and Fig. C.20 uses the 25FW outcome. Again, Figures X and Y show the analogous plots for the Cox proportional hazards model with a lasso penalty.
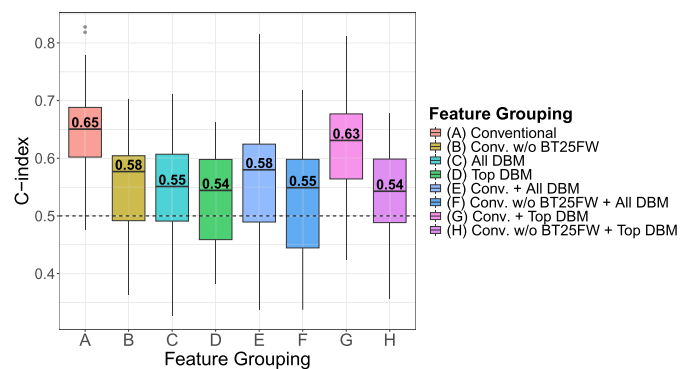
Fig. C.16. S25FW Cox model results (lasso). Boxplots show the distribution of penalized Cox proportional hazards model (lasso penalty) test-set C-indices across all cross-validation folds and repeats for outcome 25FW. The *x*-axis consists of the five feature groupings and the *y*-axis shows the test set C-indices. "Top MRI" refers to the top six DBM features. The annotated C-index represents the median value for a specific feature grouping.
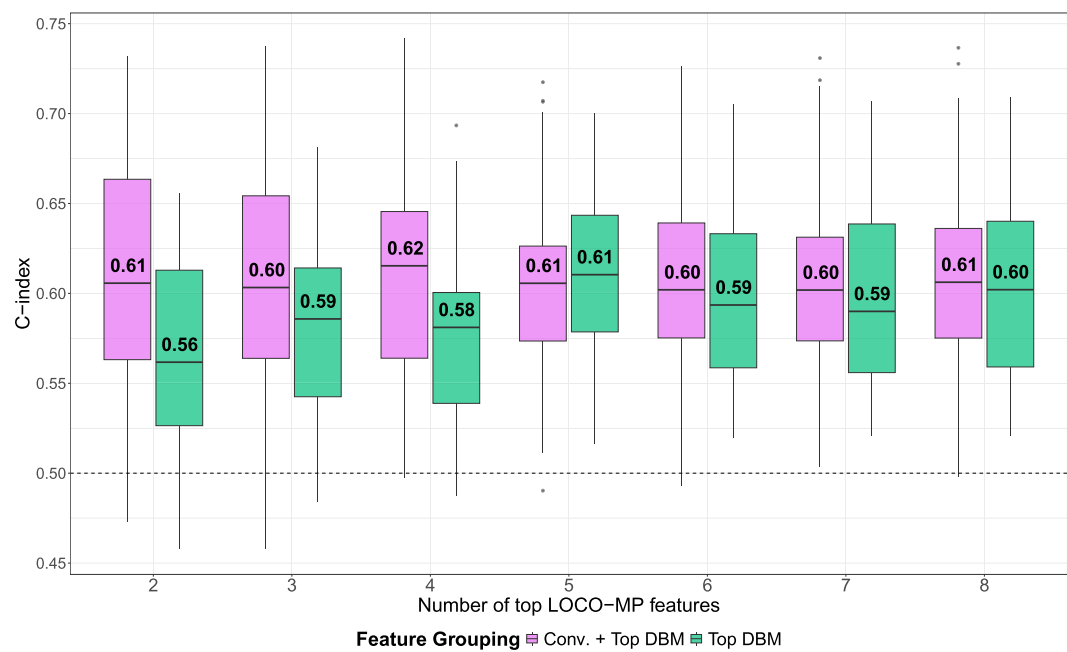

Fig. C.17. CCDP24 Cox model stability (ridge). Boxplots show the distribution of CCDP24 test-set C-indices for the conventional + top MRI and top MRI-only Cox proportional hazard (ridge penalty) models, plotted across varying numbers of top LOCO-MP features. The *x*-axis represents the number of top LOCO-MP features selected and the *y*-axis is the C-index.
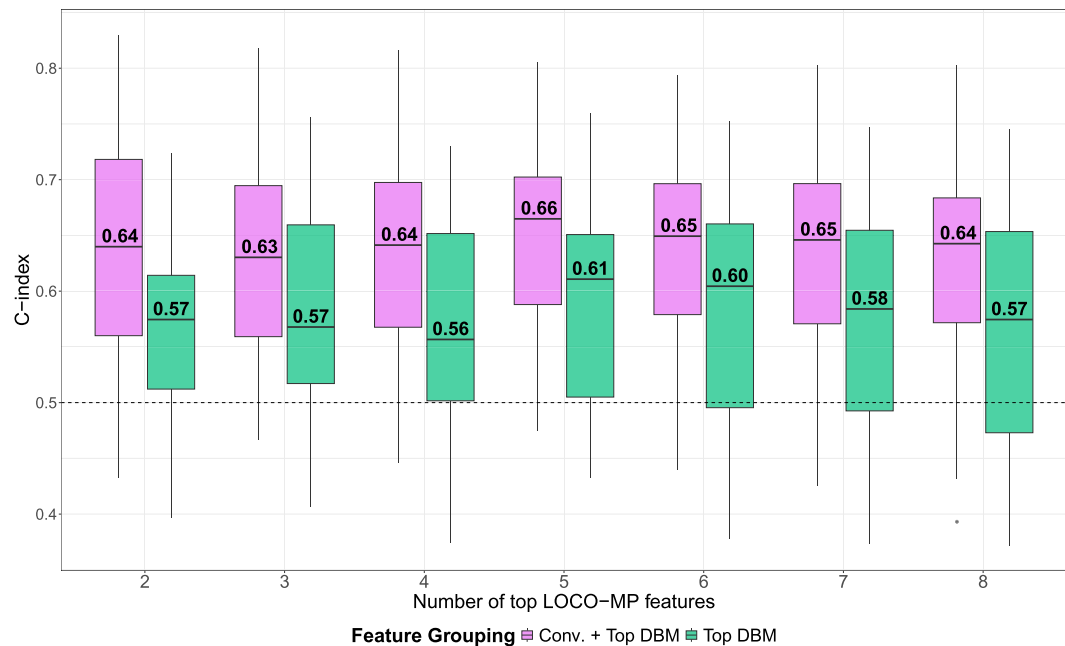
Fig. C.18. S25FW Cox model stability (ridge). Boxplots show the distribution of S25FW test-set C-indices for the conventional + top MRI and top MRI-only Cox proportional hazards (ridge penalty) models, plotted across varying numbers of top LOCO-MP features. The *x*-axis represents the number of top LOCO-MP features selected and the *y*-axis is the C-index.
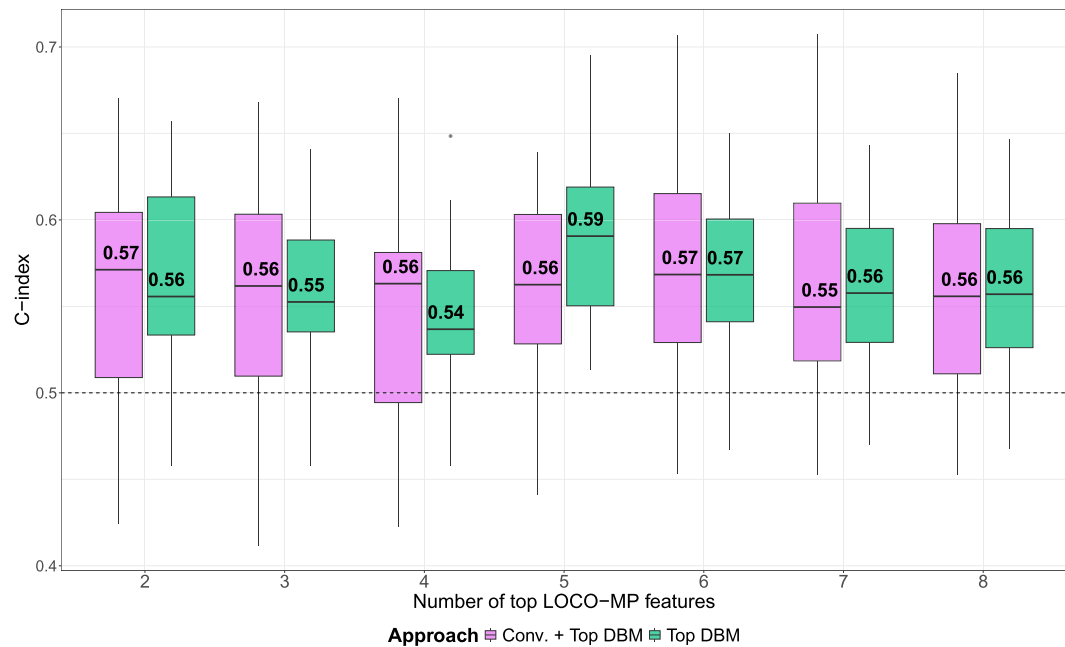

Fig. C.19. CCDP24 Cox model stability (lasso). Boxplots show the distribution of CCDP24 test-set C-indices for the conventional + top MRI and top MRI-only Cox proportional hazard (lasso penalty) models, plotted across varying numbers of top LOCO-MP features. The *x*-axis represents the number of top LOCO-MP features selected and the *y*-axis is the C-index.
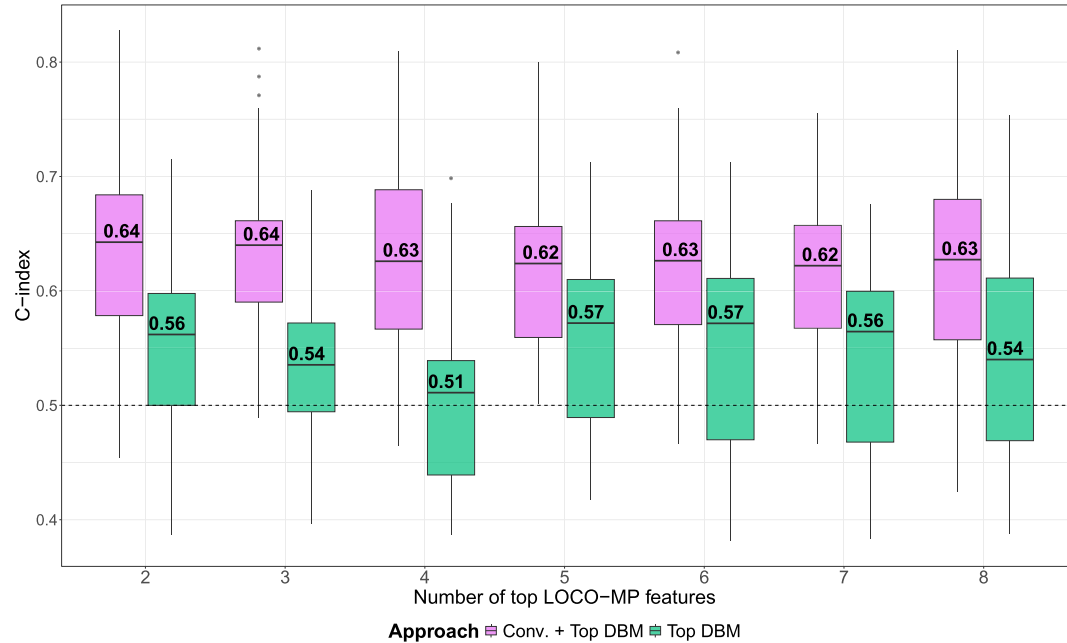
Fig. C.20. S25FW Cox model stability (lasso). Boxplots show the distribution of S25FW test-set C-indices for the conventional + top MRI and top MRI-only Cox proportional hazards (lasso penalty) models, plotted across varying numbers of top LOCO-MP features. The *x*-axis represents the number of top LOCO-MP features selected and the *y*-axis is the C-index.

Appendix D. Generalizability across study arms

To evaluate the generalizability of the LOCO-MP features, we replicated the random survival forest models in Section 3.2 using an additional internal control cohort that is demographically matched to OPERA I (OPERA II: NCT01412333) [26]. This arm was previously not used for feature selection or modeling and serves as an independent validation cohort. We use the same feature groupings and modeling process outlined in Section 2.3 to assess whether the LOCO-MP features identified with OPERA I can generalize to an unseen cohort. The RSF models are retrained on OPERA II for CCDP24 and S25FW progression with and without the conventional features, with hyperparameter tuning for number of trees. We compare models using all DBM features versus the LOCO-MP selected top DBM features, and look at median performance over thirty 80/20 data splits, analogously to the main text. Table D.6 shows that the Top DBM features induce substantial C-Index gains over all DBM features for each retrained model set.

Table D.6
Median C-Index improvements when using sparse DBM feature set over all DBM features in OPERA II validation cohort ($n = 340$). Results are from thirty 80/20 splits of the data.

Outcome	Includes Conventional Features	Median C-Index Δ of Top DBM
CCDP	No	+0.06
CCDP	Yes	+0.04
S25FW	No	+0.15
S25FW	Yes	+0.12

Data availability

The fully anonymized, individual patient raw data including clinical and MRI data of the OPERA trials are made available through the International Progressive MS Alliance (www.progressivemsalliance.org).

References

[1] Compston A, Coles A. Multiple sclerosis. Lancet 2008;372(9648):1502–17.

[2] Chard D, Griffin C, Parker G, Kapoor R, Thompson A, Miller D. Brain atrophy in clinically early relapsing–remitting multiple sclerosis. Brain 2002;125(2):327–37.

[3] Bermel RA, Bakshi R. The measurement and clinical relevance of brain atrophy in multiple sclerosis. Lancet Neurol 2006;5(2):158–70.

[4] Lomer NB, Asalemi KA, Saberi A, Sarlak K. Predictors of multiple sclerosis progression: a systematic review of conventional magnetic resonance imaging studies. PLoS ONE 2024;19(4):e0300415. https://doi.org/10.1371/journal.pone.0300415.

[5] Fisniku LK, Chard DT, Jackson JS, Anderson VM, Altmann DR, Miszkiel KA, et al. Gray matter atrophy is related to long-term disability in multiple sclerosis. Ann Neurol 2008;64(3):247–54.

[6] Song Z, Krishnan A, Gaetano L, Tustison NJ, Clayton D, de Crespigny A, et al. Deformation-based morphometry identifies deep brain structures protected by ocrelizumab. NeuroImage Clin 2022;34:102959. https://doi.org/10.1016/j.nicl.2022.102959.

[7] Ashburner J, Friston KJ. Voxel-based morphometry—the methods. NeuroImage 2000;11(6):805–21.

[8] Lansley J, Mataix-Cols D, Grau M, Radua J, Sastre-Garriga J. Localized grey matter atrophy in multiple sclerosis: a meta-analysis of voxel-based morphometry studies and associations with functional disability. Neurosci Biobehav Rev 2013;37(5):819–30.

[9] Ashburner J, Hutton C, Frackowiak R, Johnsrude I, Price C, Friston K. Identifying global anatomical differences: deformation-based morphometry. Hum Brain Mapp 1998;6(5–6):348–57.

[10] Chung MK, Worsley KJ, Paus T, Cherif C, Collins DL, Giedd JN, et al. A unified statistical approach to deformation-based morphometry. NeuroImage 2001;14(3):595–606.

[11] Manera AL, Dadar M, Collins DL, Ducharme S, Initiative FLDN. Deformation based morphometry study of longitudinal mri changes in behavioral variant frontotemporal dementia. NeuroImage Clin 2019;24:102079. https://doi.org/10.1016/j.nicl.2019.102079.

[12] Rocca MA, Preziosa P, Filippi M. Application of advanced mri techniques to monitor pharmacologic and rehabilitative treatment in multiple sclerosis: current status and future perspectives. Expert Rev Neurother 2019;19(9):835–66.

[13] Wattjes MP, Rovira À, Miller D, Yousry TA, Sormani MP, De Stefano N, et al. Magnims consensus guidelines on the use of mri in multiple sclerosis–establishing disease prognosis and monitoring patients. Nat Rev Neurol 2015;11(10):597–607.

[14] Moazami F, Lefevre-Utile A, Papaloukas C, Soumelis V. Machine learning approaches in study of multiple sclerosis disease through magnetic resonance images. Front Immunol 2021;12:700582. https://doi.org/10.3389/fimmu.2021.700582.

[15] Vázquez-Marrufo M, Sarrias-Arrabal E, García-Torres M, Martín-Clemente R, Izquierdo G. A systematic review of the application of machine-learning algorithms in multiple sclerosis. Neurología (Engl Ed) 2023;38(8):577–90.

[16] Prathapan V, Eipert P, Wigger N, Kipp M, Appali R, Schmitt O. Modeling and simulation for prediction of multiple sclerosis progression: a review and perspective. Comput Biol Med 2024:108416. https://doi.org/10.1016/j.compbiomed.2024.108416.

[17] Pisani AI, Scalfari A, Crescenzo F, Romualdi C, Calabrese M. A novel prognostic score to assess the risk of progression in relapsing- remitting multiple sclerosis patients. Eur J Neurol 2021;28(8):2503–12.

[18] Pellegrini F, Copetti M, Sormani MP, Bovis F, de Moor C, Debray TP, et al. Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. Int J MS Care 2020;26(14):1828–36.

[19] Bergsland N, Horakova D, Dwyer MG, Uher T, Vaneckova M, Tyblova M, et al. Gray matter atrophy patterns in multiple sclerosis: a 10-year source-based morphometry study. NeuroImage Clin 2018;17:444–51.

[20] Tousignant A, Lemaître P, Precup D, Arnold DL, Arbel T. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data. In: International conference on medical imaging with deep learning. PMLR; 2019. p. 483–92.

[21] Zhang K, Lincoln JA, Jiang X, Bernstam EV, Shams S. Predicting multiple sclerosis severity with multimodal deep neural networks. BMC Med Inform Decis Mak 2023;23(1):255. https://doi.org/10.1186/s12911-023-02354-6.

[22] Coll L, Pareto D, Carbonell-Mirabent P, Cobo-Calvo Á, Arrambide G, Vidal-Jordana Á, et al. Global and regional deep learning models for multiple sclerosis stratification from mri. J Magn Reson Imaging 2024;60(1):258–67.

[23] Sormani M, Bonzano L, Roccatagliata L, Mancardi G, Uccelli A, Bruzzi P. Surrogate endpoints for edss worsening in multiple sclerosis: a meta-analytic approach. Neurology 2010;75(4):302–9.

[24] Cadavid D, Cohen JA, Freedman MS, Goldman MD, Hartung H-P, Havrdova E, et al. The edss-plus, an improved endpoint for disability progression in secondary progressive multiple sclerosis. Int J MS Care 2017;23(1):94–105.

[25] Yousef H, Malagurski Tortei B, Castiglione F. Predicting multiple sclerosis disease progression and outcomes with machine learning and MRI-based biomarkers: a review. J Neurol 2024;271(10):6543–72.

[26] Hauser SL, Bar-Or A, Comi G, Giovannoni G, Hartung H-P, Hemmer B, et al. Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. N Engl J Med 2017;376(3):221–34.

[27] Lei J, Wasserman L. Distribution-free prediction bands for non-parametric regression. J R Stat Soc, Ser B Stat Methodol 2014;76(1):71–96.

[28] Gan L, Zheng L, Allen GI. Model-agnostic confidence intervals for feature importance: a fast and powerful approach using minipatch ensembles. arXiv:2206.02088, 2023.

[29] Agarwal A, Kenney AM, Tan YS, Tang TM, Yu B. Mdi+: a flexible random forest-based feature importance framework. arXiv:2307.01932, 2023.

[30] Dai Z, Liu H, Le QV, Tan M. Coatnet: marrying convolution and attention for all data sizes. Adv Neural Inf Process Syst 2021;34:3965–77.

[31] Yu X, Yang Q, Zhou Y, Cai LY, Gao R, Lee HH, et al. Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. Med Image Anal 2023;90:102939. https://doi.org/10.1016/j.media.2023.102939.

[32] Elliott C, Belachew S, Wolinsky JS, Hauser SL, Kappos L, Barkhof F, et al. Chronic white matter lesion activity predicts clinical progression in primary progressive multiple sclerosis. Brain 2019;142(9):2787–99.

[33] Francis SJ. Automatic lesion identification in mri of multiple sclerosis patients. Master's thesis. Montréal: McGill University; 2004.

[34] Bar-Or A, Thanei G-A, Harp C, Bernasconi C, Bonati U, Cross AH, et al. Blood neurofilament light levels predict non-relapsing progression following anti-cd20 therapy in relapsing and primary progressive multiple sclerosis: findings from the ocrelizumab randomised, double-blind phase 3 clinical trials. eBioMedicine 2023;93:104662. https://doi.org/10.1016/j.ebiom.2023.104662.

[35] Tustison NJ, Holbrook AJ, Avants BB, Roberts JM, Cook PA, Reagh ZM, et al. Longitudinal mapping of cortical thickness measurements: an Alzheimer's disease neuroimaging initiative-based evaluation study. J Alzheimer's Dis 2019;71(1):165–83.

[36] Klein A, Ghosh SS, Bao FS, Giard J, Häme Y, Stavsky E, et al. Mindboggling morphometry of human brains. PLoS Comput Biol 2017;13(2):e1005350. https://doi.org/10.1371/journal.pcbi.1005350.

[37] 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.

[38] Manjón JV, Coupé P, Martí-Bonmatí L, Collins DL, Robles M. Adaptive non-local means denoising of mr images with spatially varying noise levels. J Magn Reson Imaging 2010;31(1):192–203.

[39] Yao T, Allen GI. Feature selection for huge data via minipatch learning. arXiv:2010.08529, 2021.

[40] Gan L, Allen GI. Fast and interpretable consensus clustering via minipatch learning. PLoS Comput Biol 2022;18(10):e1010577. https://doi.org/10.1371/journal.pcbi.1010577.

[41] Zadeh SG, Schmid M. Bias in cross-entropy-based training of deep survival networks. IEEE Trans Pattern Anal Mach Intell 2020;43(9):3126–37.

[42] Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. Sci Rep 2021;11(1):13505. https://doi.org/10.1038/s41598-021-92799-4.

[43] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2(3):841–60.

[44] Ishwaran H, Kogalur UB. Fast unified random forests for survival, regression, and classification (rf-src). R package version 2 (1), Berlin: Springer; 2019.

[45] Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, et al. Package 'glmnet'. CRAN R Repositary 595, 2021.

[46] Storelli L, Azzimonti M, Gueye M, Vizzino C, Preziosa P, Tedeschi G, et al. A deep learning approach to predicting disease progression in multiple sclerosis using magnetic resonance imaging. Invest Radiol 2022;57(7):423–32.

[47] Solovyev R, Kalinin AA, Gabruseva T. 3d convolutional neural networks for stalled brain capillary detection. Comput Biol Med 2022;141:105089. https://doi.org/10.1016/j.compbiomed.2021.105089.

[48] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–55.

[49] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems, vol. 30. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[50] Yu B. Veridical data science. In: Proceedings of the 13th international conference on web search and data mining; 2020. p. 4–5.

[51] Yu B, Barter RL. Veridical data science: the practice of responsible data analysis and decision making. MIT Press; 2024.

[52] Müller M, Esser R, Kötter K, Voss J, Müller A, Stellmes P. Third ventricular enlargement in early stages of multiple sclerosis is a predictor of motor and neuropsychological deficits: a cross-sectional study. BMJ Open 2013;3(9):e003582. https://doi.org/10.1136/bmjopen-2013-003582.

[53] Guenter W, Betscher E, Bonek R. Predictive value of the third ventricle width for neurological status in multiple sclerosis. J Clin Med 2022;11(10):2841. https://doi.org/10.3390/jcm11102841.

[54] Simon J, Jacobs L, Campion M, Rudick R, Cookfair D, Herndon R, et al. A longitudinal study of brain atrophy in relapsing multiple sclerosis. Neurology 1999;53(1):139–48.

[55] Weier K, Banwell B, Cerasa A, Collins DL, Dogonowski A-M, Lassmann H, et al. The role of the cerebellum in multiple sclerosis. Cerebellum 2015;14:364–74.

[56] Brouwer EJ, Strik M, Schoonheim MM. The role of the cerebellum in multiple sclerosis: structural damage and disconnecting networks. Curr Opin Behav Sci 2024;59:101436. https://doi.org/10.1016/j.cobeha.2024.101436.

[57] Gera G, Fling BW, Horak FB. Cerebellar white matter damage is associated with postural sway deficits in people with multiple sclerosis. Arch Phys Med Rehabil 2020;101(2):258–64.

[58] Chen MH, Wylie GR, Sandroff BM, Dacosta-Aguayo R, DeLuca J, Genova HM. Neural mechanisms underlying state mental fatigue in multiple sclerosis: a pilot study. J Neurol 2020;267:2372–82.

[59] Hulst HE, Schoonheim MM, Roosendaal SD, Popescu V, Schweren LJ, van der Werf YD, et al. Functional adaptive changes within the hippocampal memory system of patients with multiple sclerosis. Hum Brain Mapp 2012;33(10):2268–80.

[60] Geurts JJ, Bö L, Roosendaal SD, Hazes T, Daniëls R, Barkhof F, et al. Extensive hippocampal demyelination in multiple sclerosis. J Neuropathol Exp Neurol 2007;66(9):819–27.

[61] Eshaghi A, Marinescu RV, Young AL, Firth NC, Prados F, Jorge Cardoso M, et al. Progression of regional grey matter atrophy in multiple sclerosis. Brain 2018;141(6):1665–77.

[62] Montalban X, Hauser SL, Kappos L, Arnold DL, Bar-Or A, Comi G, et al. Ocrelizumab versus placebo in primary progressive multiple sclerosis. N Engl J Med 2017;376(3):209–20.

[63] Krishnan AP, Song Z, Clayton D, Jia X, de Crespigny A, Carano RA. Multi-arm u-net with dense input and skip connectivity for t2 lesion segmentation in clinical trials of multiple sclerosis. Sci Rep 2023;13(1):4102. https://doi.org/10.1038/s41598-023-31207-5.

[64] Krajnc N, Berger T, Bsteh G. Measuring treatment response in progressive multiple sclerosis—considerations for adapting to an era of multiple treatment options. Biomolecules 2021;11(9):1342. https://doi.org/10.3390/biom11091342.

[65] Sikes EM, Cederberg KL, Sandroff BM, Bartolucci A, Motl RW. Quantitative synthesis of timed 25-foot walk performance in multiple sclerosis. Arch Phys Med Rehabil 2020;101(3):524–34.

[66] Koch MW, Mostert J, Uitdehaag B, Cutter G. Clinical outcome measures in spms trials: an analysis of the impact and ascend original trial data sets. Mult Scler J 2020;26(12):1540–9.

[67] Branco R, Valente JB, Martins AS, Soares DF, Castanho EN, Madeira SC, et al. Survival analysis for multiple sclerosis: predicting risk of disease worsening. In: CLEF (working notes); 2023. p. 1199–209.

[68] Guazzo A, Trescato I, Longato E, Tavazzi E, Vettoretti M, Di Camillo B, et al. Baseline machine learning approaches to predict multiple sclerosis disease progression. In: CLEF (working notes); 2023. p. 1219–32.

[69] Andorra M, Freire A, Zubizarreta I, de Rosbo NK, Bos SD, Rinas M, et al. Predicting disease severity in multiple sclerosis using multimodal data and machine learning. J Neurol 2024;271(3):1133–49.

[70] Fuh-Ngwa V, Zhou Y, Charlesworth JC, Ponsonby A-L, Simpson-Yap S, et al. Developing a clinical–environmental–genotypic prognostic index for relapsing-onset multiple sclerosis and clinically isolated syndrome. Brain Commun 2021;3(4):fcab288. https://doi.org/10.1093/braincomms/fcab288.

[71] Hossain MZ, Daskalaki E, Brüstle A, Desborough J, Lueck CJ, Suominen H. The role of machine learning in developing non-magnetic resonance imaging based biomarkers for multiple sclerosis: a systematic review. BMC Med Inform Decis Mak 2022;22(1):242.

[72] Cohen M, Bresch S, Rocchi OT, Morain E, Benoit J, et al. Should we still only rely on edss to evaluate disability in multiple sclerosis patients? A study of inter and intra rater reliability. Mult Scler Relat Disord 2021;54:103144. https://doi.org/10.1016/j.msard.2021.103144.

[73] Motl RW, Cohen JA, Consortium MSOA. Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis. Mult Scler J 2017;23(5):704–10. https://doi.org/10.1177/1352458517690823.

[74] Solaro C, Cattaneo D, Brichetto G, Tacchino A, Gervasoni E, et al. Clinical correlates of 9-hole peg test in a large population of people with multiple sclerosis. Mult Scler Relat Disord 2019;30:1–8. https://doi.org/10.1016/j.msard.2019.01.043.

[75] Preziosa P, Pagani E, Meani A, Moiola L, Rodegher M, Filippi M, et al. Slowly expanding lesions predict 9-year multiple sclerosis disease progression. Neurology: Neuroimmunol Neuroinflamm 2022;9(2):e1139. https://doi.org/10.1212/NXI.0000000000001139.

[76] Shao Z, Chen Y, Bian H, Zhang J, Liu G, Zhang Y. HVTSurv: hierarchical vision transformer for patient-level survival prediction from whole slide image. In: Proceedings of the AAAI conference on artificial intelligence, vol. 37; 2023. p. 2209–17.

[77] Ridgeway G, Developers G. gbm: generalized boosted regression models. r package version 2.2.2, https://CRAN.R-project.org/package=gbm, 2024.