MSJ Experimental Translational Clinical

Original Research Article

Validation of a machine learning approach to estimate expanded disability status scale scores for multiple sclerosis

Multiple Sclerosis Journal—Experimental, Translational and Clinical

April–June 2022, 1–8

DOI: 10.1177/20552173221108635

Pedro Alves, Eric Green, Michelle Leavy, Haley Friedler, Gary Curhan, Carl Marci and Costas Boussios

Abstract

Background: Disability assessment using the Expanded Disability Status Scale (EDSS) is important to inform treatment decisions and monitor the progression of multiple sclerosis. Yet, EDSS scores are documented infrequently in electronic medical records.

Objective: To validate a machine learning model to estimate EDSS scores for multiple sclerosis patients using clinical notes from neurologists.

Methods: A machine learning model was developed to estimate EDSS scores on specific encounter dates using clinical notes from neurologist visits. The OM1 MS Registry data were used to create a training cohort of 2632 encounters and a separate validation cohort of 857 encounters, all with clinician-recorded EDSS scores. Model performance was assessed using the area under the receiver-operating-characteristic curve (AUC), positive predictive value (PPV), and negative predictive value (NPV), calculated using a binarized version of the outcome. The Spearman *R* and Pearson *R* values were calculated. The model was then applied to encounters without clinician-recorded EDSS scores in the MS Registry.

Results: The model had a PPV of 0.85, NPV of 0.85, and AUC of 0.91. The model had a Spearman *R* value of 0.75 and Pearson *R* value of 0.74 when evaluating performance using the continuous estimated EDSS and clinician-recorded EDSS scores. Application of the model to eligible encounters resulted in the generation of eEDSS scores for an additional 190,282 encounters from 13,249 patients.

Conclusion: EDSS scores can be estimated with very good performance using a machine learning model applied to clinical notes, thus increasing the utility of real-world data sources for research purposes.

Keywords: Multiple sclerosis, health services research, machine learning, disability evaluation, outcome assessment, registries

Date received: 3 January 2022; accepted 6 June 2022

Introduction

Assessment of multiple sclerosis (MS) disease activity is critical for understanding treatment effectiveness, informing treatment decisions, and monitoring disease progression. Magnetic resonance imaging (MRI) findings, clinician-administered assessments, and patient questionnaires are used to measure disease activity and track changes over time. The Expanded Disability Status Scale (EDSS) is a validated, clinician-administered scale used to measure disability in MS.[1] The scale considers the impairment in the function of the pyramidal, cerebellar, brainstem, sensory, bladder, bowel, visual, and cerebral systems.

The EDSS is widely used in clinical trials,[2] but its use in routine clinical practice is limited due to the time required to complete the scale and the complexity of scoring.[3] As a result, EDSS scores are documented rarely and inconsistently in real-world data sources such as electronic medical records (EMRs). Real-world data sources have played an increasingly important role in MS research in recent years,[4] but

Correspondence to:
Michelle Leavy,
Research, OM1 Inc., 31 St. James Ave, Boston, MA 02116, USA.
mleavy@om1.com

Pedro Alves,
Eric Green,
Data Science, OM1, Inc., Boston, MA, USA

Michelle Leavy,
Haley Friedler,
Gary Curhan,
Research, OM1, Inc., Boston, MA, USA

Carl Marci,
Mental Health and
Neuroscience, OM1, Inc.,
Boston, MA, USA

Costas Boussios,
Data Science, OM1, Inc.,
Boston, MA, USA

the lack of EDSS scores in these data sources makes it difficult to address some questions related to disease progression, treatment patterns, and patient outcomes. While existing statistical methods for imputing missing scores can be used in some research studies, many patients in real-world data sets lack the necessary EDSS scores over time to make imputation feasible.

Machine learning offers a new approach to addressing the lack of disease activity scores captured in real-world data sources.[5,6] Machine learning models have been developed to predict MS relapse risk,[7] disease progression,[8] disability progression,[9] and EDSS scores at a future timepoints using clinical and other data.[10–12] These efforts highlight the potential of machine learning approaches in MS and demonstrate that it is feasible to predict EDSS scores using routinely recorded clinical data. While prediction of future EDSS scores is useful for informing treatment decisions, estimation of EDSS scores at discrete timepoints throughout the course of illness would increase the value of real-world data sources for MS research.

The objective of this study was to validate a machine learning model to generate estimated EDSS (eEDSS) scores at specific timepoints for MS patients using available clinical notes from a real-world data source.

Patients and methods

Participants

Data for this study were drawn from the OM1 MS Registry (OM1, Inc, Boston, MA). The OM1 MS Registry is a commercially available syndicated dataset derived from the OM1 Real-World Data Cloud, a multisource database of deterministically linked, de-identified, patient-level health care information including EMR clinical notes, unadjudicated insurance, and pharmacy claims, and other data from 2013 to 2021 for specialty and primary care patients across the United States. Patients from the OM1 Real-World Data Cloud are eligible for inclusion in the OM1 MS Registry if they are followed by a neurologist and meet one of the following criteria: (1) two or more diagnosis codes for MS; (2) at least one demyelinating disorder diagnosis code followed by an MS diagnosis code; or (3) one MS diagnosis code and a prescription for an MS-related disease-modifying therapy. The OM1 MS Registry includes clinical and administrative data on over 19,000 MS patients who meet one or more of these criteria. Registry data include clinical data extracted from EMRs (e.g. medication and prescription history, laboratory results, and diagnoses), unstructured physician-documented notes, and unadjudicated medical and pharmacy claims data. Clinician-recorded EDSS scores are extracted from clinical notes where available and are included in the registry. Registry data are de-identified. The study was submitted for Institutional Review Board approval and determined to be exempt.

For this study, we identified encounters in the MS Registry with a text-based clinical note as well as a clinician-recorded EDSS score. We restricted this analysis to the largest data source. Within that data source, we identified four key clinical evaluation sections (history of present illness, physical exam, review of systems, and clinical assessment) in the unstructured portion of the notes. Only notes that had at least one of these four clinical evaluation sections with sufficient clinical detail to generate an eEDSS score were included. These clinical evaluation sections contain information related to the evolution and severity of MS and, as such, were deemed critical for generating the estimation model. We then randomly assigned patients to either the model training cohort (75%) or the model validation cohort (25%). Finally, we created the training data set using all encounters from patients assigned to the training cohort, and we created the validation data set using all encounters from patients assigned to the validation cohort.

Modeling strategy

Dependent variable. The EDSS is an ordinal scale, with scores ranging from 0 to 10 in 0.5 increments. A score of 0 indicates normal neurological status, while a score of 10 indicates death due to MS. Scores of 5 or higher reflect ambulatory impairment, with scores of 6 and higher indicating the need for ambulatory aid. Many studies have reported a bimodal frequency distribution in EDSS scores, with scores of 3 and 6 occurring most frequently.[2] Using the training cohort, the model is trained to produce estimated scores ranging from 1 to 10 in increments of 0.5. The trained model is then used to generate an eEDSS score for a specific encounter on a specific date for patients in the validation cohort.

Explanatory variables. Model features were derived from the clinical evaluation sections of the unstructured portion of the clinical notes. The explanatory terms and phrases included those indicating mobility impairments; presence or negation of symptoms such as loss of balance, speech impairment, vision impairment, muscle spasms, and others; explicit indications of disease progression such as improvement or deterioration of symptoms; and mentions of medications (see Supplemental Table 1 for additional examples of explanatory terms and phrases).

Model development. The algorithm processes the body of clinical notes to de-noise and standardize their content and subsequently to derive the set of explanatory features described above. Covariate influence is measured using the permutation importance methodology.[13] The EDSS estimation model uses a pair of XGBoost gradient boosting regression models accompanied by a binary decision node which specifies which regression model to be used. The decision node processes the mobility impairment features in the medical note and determines whether the patient requires mobility aid. Additional details relating to developing a model from clinical notes are included in a separate publication.[5]

Model performance. We calculated the Spearman *R* and Pearson *R* values to evaluate the eEDSS scores versus the clinician-recorded EDSS scores on a continuous scale. We also assessed the performance of the model as a binary predictor using the area under the receiver-operating-characteristic curve (AUC), positive predictive value (PPV), and negative predictive value (NPV). We used a binarized version of the outcome in which the positive class is defined as those notes with scores greater or equal to 6 (the threshold at which EDSS scores reflects the need for ambulatory aid), and the negative class is defined as those records with scores less than or equal to 5.5. The cutoff of 6 was selected due to the clinical relevance of the score.[2,14] Lastly, binary performance metrics for cutoffs other than 6 and agreement percentages of eEDSS and clinician-recorded EDSS for different score ranges were calculated to further assess the discriminatory capabilities of the model.

As a final step, we reviewed the model features for clinical suitability and compared the distribution of eEDSS scores to the distribution of clinician-recorded EDSS scores. The model was then applied to encounter notes that had the required clinical evaluation sections but no clinician-recorded EDSS scores in the MS Registry.

Results

Participants and characteristics

At the time of model development, the data source used for this study included 13,766 patients, 684 of

Table 1. Demographic and clinical characteristics of training and validation cohorts.

		Training cohort ($n = 513$)	Validation cohort ($n = 171$)
Age, yrs	Mean (s.d.)	53.5 (12.6)	52.1 (13.7)
Sex	Female	388 (75.6%)	125 (73.1%)
	Male	125 (24.4%)	46 (26.9%)
Race	White	404 (78.8%)	138 (81.7%)
	Black	44 (8.6%)	13 (7.7%)
	Other	65 (12.7%)	18 (10.7%)
	Unknown	0	2
Region	Northeast	16 (3.4%)	2 (1.3%)
	Midwest	242 (51.8%)	83 (53.9%)
	South	18 (3.9%)	7 (4.5%)
	West	191 (40.9%)	62 (40.3%)
	Unknown	46	17
Duration of follow-up, yrs	Mean (s.d.)	4.5 (2.2)	4.6 (2.1)
Number of encounters	Mean (s.d.)	5.1 (3.4)	5.0 (3.3)
Clinician-recorded EDSS	Mean (s.d.)	4.0 (2.3)	4.1 (2.3)
DMT use	*n* (%)	369 (71.9%)	136 (79.5%)
Corticosteroid use	*n* (%)	213 (41.5%)	69 (40.4%)
Depression	*n* (%)	112 (21.8%)	35 (20.5%)
Anxiety	*n* (%)	124 (24.2%)	42 (24.6%)
Hypertension	*n* (%)	258 (50.3%)	82 (48.0%)

DMT: disease-modifying therapies; EDSS: Expanded Disability Status Scale Score.

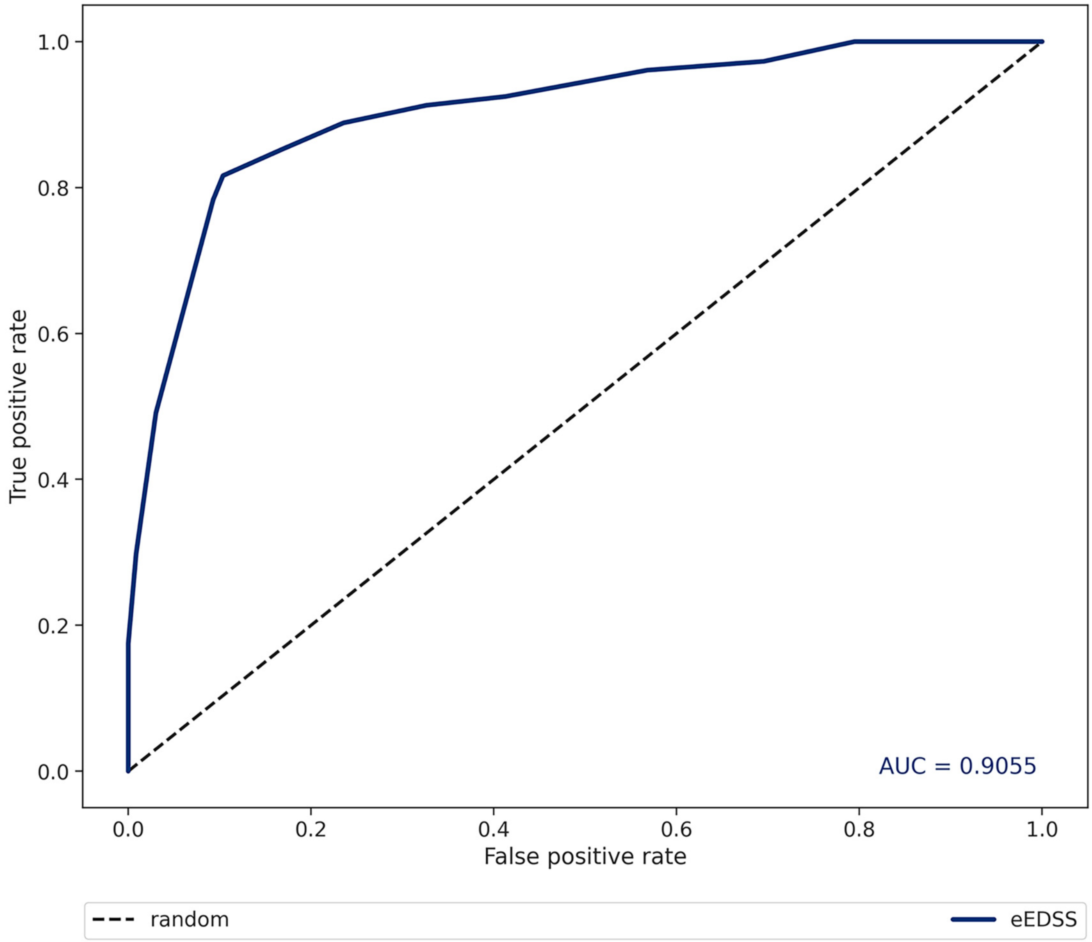


Figure 1. The area under the receiver-operating-characteristic curve (AUC). The AUC was calculated using a binarized version of the outcome in which the positive class is defined as those notes with scores greater or equal to 6 (the threshold at which EDSS scores reflects the requirement for ambulatory aid), and the negative class is defined as those records with scores less than or equal to 5.5.
AUC: area under the receiver-operating-characteristic curve; EDSS: Expanded Disability Status Scale score.

whom had an EDSS score and clinical notes from the encounter. The model training cohort consisted of 2632 encounters from 513 patients, while the validation cohort consisted of 857 encounters from 171 patients. Table 1 presents the demographic and clinical characteristics of the patients in the training cohort and validation cohort. Demographics, patient characteristics, and common comorbidities are similar between the training cohort and the validation cohort.

Model performance

The model had a Spearman *R* value of 0.75 and Pearson *R* value of 0.74 when evaluating performance using the continuous eEDSS and clinician-recorded EDSS scores. The model had a PPV of 0.85, NPV of 0.85, and AUC of 0.91 when evaluating performance using the binarized version of the outcome in the validation cohort (Figure 1). Binary performance metrics for cutoffs other than 6 and agreement percentages of eEDSS and clinician-recorded EDSS for different score ranges are presented in Supplemental Tables 2 and 3, respectively.

The distribution of eEDSS scores was compared to the distribution of clinician-recorded EDSS scores (Figure 2), and a confusion matrix was generated to further assess the agreement between the eEDSS scores and clinician-recorded EDSS scores (Figure 3). These figures show that the agreement between the eEDSS and clinician-recorded EDSS scores is higher for the higher scores.

As a final step, the model was applied to encounter notes without clinician-recorded EDSS scores in the data source that was used for model development. As of the end of 2021, 3489 encounters from 684 patients

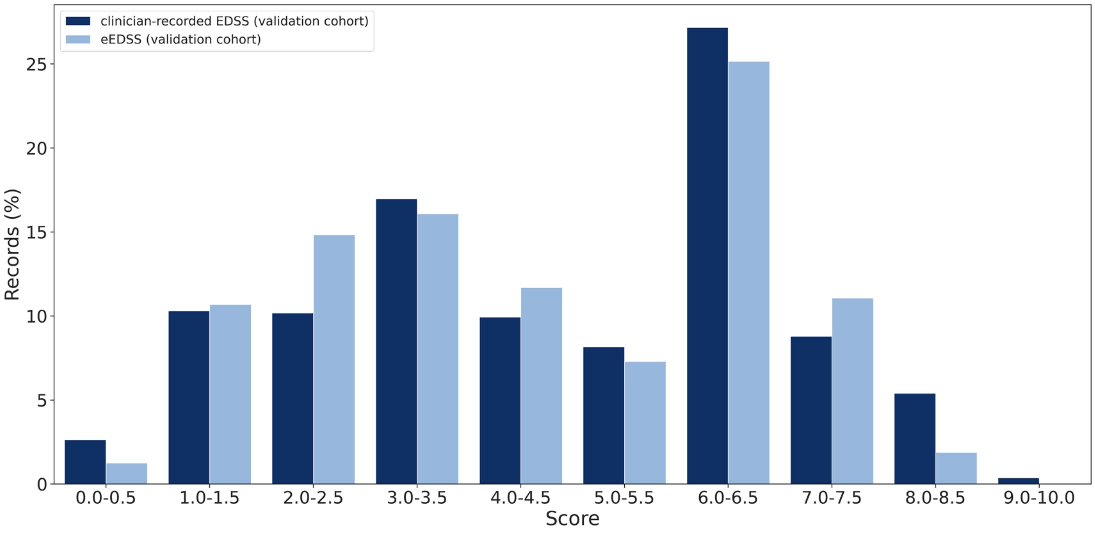


Figure 2. Distribution of estimated and clinician-recorded EDSS scores in the validation cohort. The distribution of eEDSS scores was compared to the distribution of clinician-recorded EDSS scores.
EDSS: Expanded Disability Status Scale; eEDSS: estimated EDSS.

had both a clinician-recorded EDSS score and a clinical note associated with the encounter. Application of the model to encounters that had clinical evaluation sections but no clinician-recorded EDSS scores resulted in the generation of eEDSS scores for an additional 190,282 encounters from 13,249 patients. The number of patients with at least one eEDSS score is 19.4 times as many as the number of patients with at least one clinician-recorded EDSS score only.

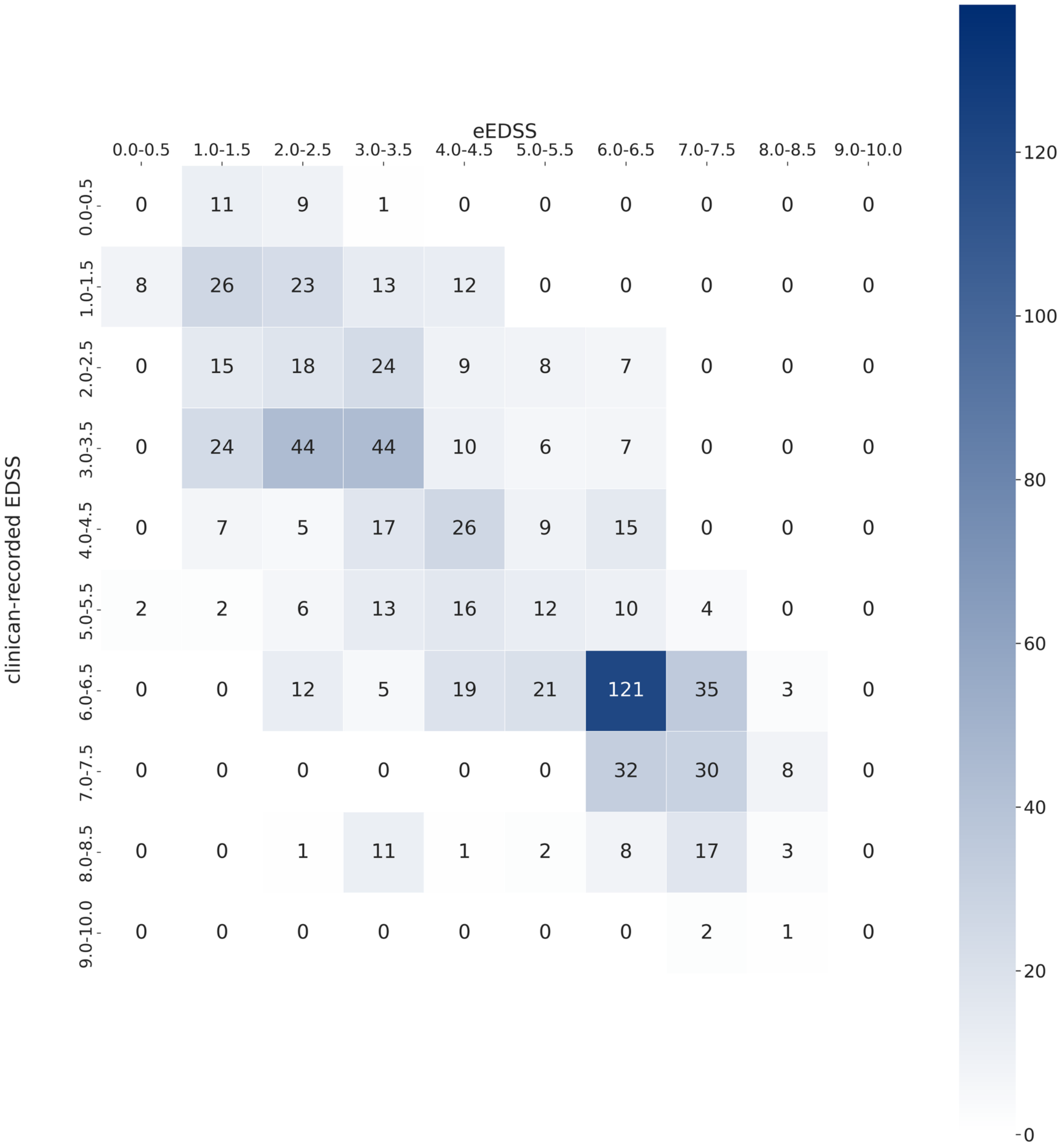
Estimated scores were not generated for 994 encounters from 222 patients because these encounters did not have at least one note section from one of the required categories, and an additional 43,219 encounters from 7819 patients were excluded due to insufficient clinical detail in the note sections. Of the 190,282 encounters for which eEDSS scores were generated, 189,235 (99.4%) had assessment notes, 188,492 (99.1%) had history of present illness notes, 176,688 (92.9%) had physical exam notes, and 173,756 (91.3%) had review of systems notes. Overall, 98.2% of encounters for which an eEDSS was generated had at least three of these sections, thus providing rich clinical content to generate the eEDSS.

The distribution of the eEDSS scores for the 190,282 encounters as compared to encounters with clinician-recorded EDSS scores is presented in Figure 4. Both distributions are bimodal, and the distribution of the eEDSS scores includes a larger proportion of lower eEDSS scores than the distribution of clinician-recorded EDSS scores.

Discussion

Measurement of disability progression is of critical importance in MS, with relevance to patients, clinicians making treatment decisions, and researchers seeking to understand treatment effectiveness, disease burden, and patient outcomes over time. The EDSS is widely used in clinical trials to measure disability progression and response to treatment, but it is rarely captured and documented in routine clinical practice. This leads to gaps in real-world data sources and limits their value as a source of insight for evaluating treatment successes and failures. This study demonstrates that a machine learning model can use clinical notes to generate an eEDSS score for specific patient encounters with very good performance.

Machine learning has been the subject of much interest in MS research in recent years, with many efforts to develop models to aid in MS diagnosis and treatment.[7–12] Our effort appears to be unique in that the machine learning model estimates a numeric EDSS score at a specific encounter date based on clinician documentation, with the goal of filling in gaps in longitudinal data from real-world data sources. Use of eEDSS scores for encounters where no clinician-recorded EDSS scores are available increases the number of scores available for retrospective analysis, thereby providing a more comprehensive view of the progression of disability over time. The model estimation of EDSS scores also enables comparison of disability and disease progression in real-world settings to disability and disease



	0.0-0.5	1.0-1.5	2.0-2.5	3.0-3.5	4.0-4.5	5.0-5.5	6.0-6.5	7.0-7.5	8.0-8.5	9.0-10.0
0.0-0.5	0	11	9	1	0	0	0	0	0	0
1.0-1.5	8	26	23	13	12	0	0	0	0	0
2.0-2.5	0	15	18	24	9	8	7	0	0	0
3.0-3.5	0	24	44	44	10	6	7	0	0	0
4.0-4.5	0	7	5	17	26	9	15	0	0	0
5.0-5.5	2	2	6	13	16	12	10	4	0	0
6.0-6.5	0	0	12	5	19	21	121	35	3	0
7.0-7.5	0	0	0	0	0	0	32	30	8	0
8.0-8.5	0	0	1	11	1	2	8	17	3	0
9.0-10.0	0	0	0	0	0	0	0	2	1	0

Figure 3. Confusion matrix showing agreement between estimated and clinician-recorded EDSS scores in the validation cohort. A confusion matrix was generated to further assess the agreement between the eEDSS scores and clinician-recorded EDSS scores.
EDSS: Expanded Disability Status Scale; eEDSS: estimated EDSS.

progression observed in clinical trials where EDSS scores are recorded by investigators at regular intervals.

The present methodology has several strengths. The model estimation we used does not aim to replicate the typical EDSS calculation by extracting each score component and computing the sum of their weights. Rather, our approach uses information from up to four relevant sections of the unstructured clinical notes (history of present illness, physical exam, review of systems, and clinical assessment) as the input to the model training and scoring. This rich, clinically relevant content allowed us to include features such as medication use, improvement or deterioration of disease, and assessments of pain that are not included in the traditional EDSS computation. The availability of a large number of clinical notes with a similar data architecture in the MS Registry further allowed us to apply the model to an additional 190,282 encounters from 13,249 patients. This level of amplification can only be accomplished with modern computational power. Another strength

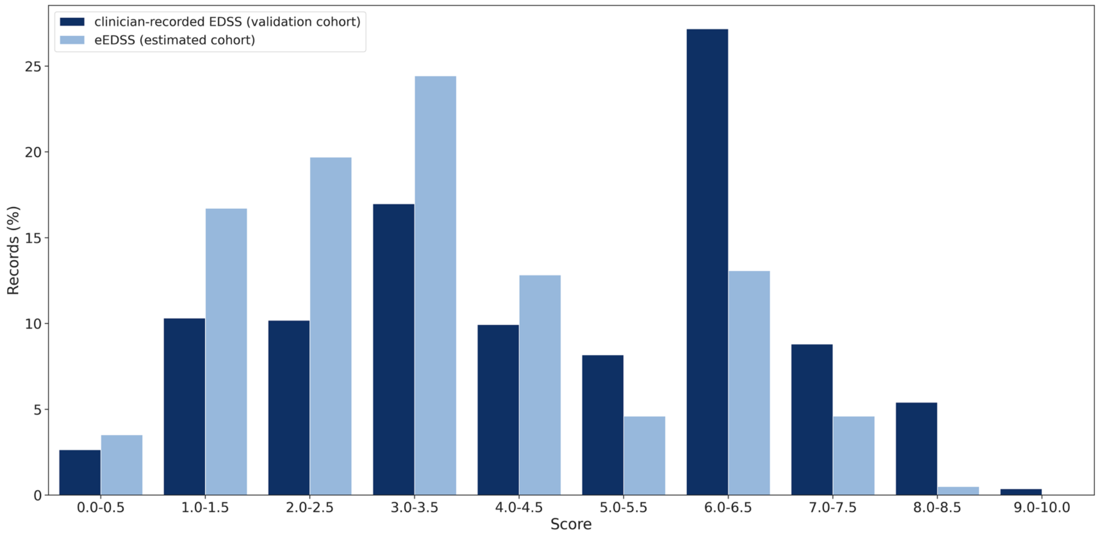


Figure 4. Distribution of estimated and clinician-recorded EDSS scores in the validation cohort. The distribution of eEDSS scores for eligible encounters in the MS Registry was compared to the distribution of clinician-recorded EDSS scores in the validation cohort, with the scores on the *x*-axis and the percentage of total encounters on the *y*-axis.
EDSS: Expanded Disability Status Scale; eEDSS: estimated EDSS.

of this effort is our interdisciplinary approach. We included data scientists, clinical experts, clinical informaticists, and software engineers in the process to generate the datasets and develop and test the model.

The machine learning model has some limitations. First, the model uses information extracted from clinical notes, and information related to disability status that is not documented in the notes is not considered in the estimation. This has the potential to lower the reliability of the estimate. However, the clinician-recorded EDSS has well-documented limitations, including suboptimal inter-rater reliability for both individual functional system sub-scores and the total EDSS scores.[15] Lower inter-rater reliability may relate in part to the subjectivity of interpreting the neurologic exam and the experience of the rater.[16] Both intra-rater and inter-rater reliability are higher at higher ranges of the EDSS scale (e.g. 6.0 and above) than in lower ranges (e.g. 1.0–3.5) where symptoms and subtle signs can affect the total score.[16] This reduced reliability is most likely present in our data source, meaning that the clinician-recorded EDSS is likely less reliable at lower scores. Conversely, the higher reliability of clinician-recorded EDSS scores for higher scores, as documented in the literature, aligns with the higher observed performance of the eEDSS for higher scores.

Second, the clinical notes used to train and validate the model are drawn from neurology practices in the United States and may not be reflective of documentation practices in other geographic locations or care settings. EMR systems and their configurations vary widely across different medical providers and specialists, and the use of medical terminology also varies broadly among practitioners. Our approach recognizes common language patterns and maps expressions with the same meaning to the same term, increasing the generalizability within the same data source. However, the model likely would have less success standardizing the note contents from a data source that is different from the training data source. Improving the model generalizability to multiple data sources that are not available during the model development phase is an important component of future research.

Disability progression is a critical outcome in MS, and the lack of clinician-recorded EDSS scores in real-world data sources decreases the value of these data sources for supporting research on MS treatments and outcomes. This study addresses this issue by developing and validating a model to generate eEDSS scores using clinical notes. Applying the model to real-world data sources has the potential to make these data sources more valuable for supporting research on MS disability and disease progression, monitoring treatment effectiveness, and improving patient outcomes.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The authors indicated are employees

of OM1, which is involved in issues related to the topic of this manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Michelle Leavy https://orcid.org/0000-0003-1927-7248
Haley Friedler https://orcid.org/0000-0003-4383-100X

Supplemental material

Supplemental material for this article is available online.

References

1. Kurtzke JF. A new scale for evaluating disability in multiple sclerosis. *Neurology* 1955; 5: 580.
2. Meyer-Moock S, Feng Y-S, Maeurer M, et al. Systematic literature review and validity evaluation of the expanded disability status scale (EDSS) and the multiple sclerosis functional composite (MSFC) in patients with multiple sclerosis. *BMC Neurol* 2014; 14: 58.
3. Baldassari LE, Salter AR, Longbrake EE, et al. Streamlined EDSS for use in multiple sclerosis clinical practice: development and cross-sectional comparison to EDSS. *Mult Scler* 2018; 24: 1347–1355.
4. Cohen JA, Trojano M, Mowry EM, et al. Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. *Mult Scler* 2020; 26: 23–37.
5. Alves P, Bandaria J, Leavy MB, et al. Validation of a machine learning approach to estimate systemic lupus erythematosus disease activity index score categories and application in a real-world dataset. *RMD Open* 2021; 7: e001586.
6. Spencer AK, Bandaria J, Leavy MB, et al. Validation of a machine learning approach to estimate clinical disease activity index scores for rheumatoid arthritis. *RMD Open* 2021; 7(3):e001781. DOI: 10.1136/rmdopen-2021-001781.
7. Ahuja Y, Kim N, Liang L, et al. Leveraging electronic health records data to predict multiple sclerosis disease activity. *Ann Clin Transl Neurol* 2021; 8: 800–810.
8. Pinto MF, Oliveira H, Batista S, et al. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Sci Rep* 2020; 10: 21038.
9. Stafford IS, Kellermann M, Mossotto E, et al. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med* 2020; 3: 30.
10. De Brouwer E, Becker T, Moreau Y, et al. Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression. *Comput Methods Programs Biomed* 2021; 208: 106180.
11. Tommasin S, Cocozza S, Taloni A, et al. Machine learning classifier to identify clinical and radiological features relevant to disability progression in multiple sclerosis. *J Neurol* 2021; 268: 4834–4845.
12. Roca P, Attye A, Colas L, et al. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagn Interv Imaging* 2020; 101: 795–802.
13. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
14. Uitdehaag BMJ. Disability outcome measures in phase III clinical trials in multiple sclerosis. *CNS Drugs* 2018; 32: 543–558.
15. Amato MP, Fratiglioni L, Groppi C, et al. Interrater reliability in assessing functional systems and disability on the Kurtzke scale in multiple sclerosis. *Arch Neurol* 1988; 45: 746–748.
16. Noseworthy JH, Vandervoort MK, Wong CJ, et al. Interrater variability with the expanded disability status scale (EDSS) and functional systems (FS) in a multiple sclerosis clinical trial. The Canadian cooperation MS study group. *Neurology* 1990; 40: 971–975.