ANNALS of Clinical and Translational Neurology
Open Access
ANA AMERICAN NEUROLOGICAL ASSOCIATION INNOVATORS IN DISCOVERY, EDUCATION, AND CARE

RESEARCH ARTICLE

Leveraging electronic health records data to predict multiple sclerosis disease activity

Yuri Ahuja[1,a], Nicole Kim[1,a], Liang Liang[1], Tianrun Cai[2], Kumar Dahal[2], Thany Seyok[2], Chen Lin[3], Sean Finan[3], Katherine Liao[2], Guergana Savovoa[3], Tanuja Chitnis[4], Tianxi Cai[1,5,b] & Zongqi Xia[6,b]

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA
[2]Division of Rheumatology, Department of Medicine, Brigham and Women's Hospital, Boston, MA
[3]Clinical Natural Language Processing Program, Boston Children's Hospital, Boston, MA
[4]Department of Neurology, Brigham and Women's Hospital, Boston, MA
[5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA
[6]Department of Neurology and Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Correspondence
Zongqi Xia, Department of Neurology, University of Pittsburgh, Biomedical Science Tower 3, Suite 7014, 3501 5th Avenue, Pittsburgh, PA 15260. Tel: +1 412-383-5377; Fax: +1 412-648-7233; E-mail: zxia1@post.harvard.edu

Funding Information
National Institute of Neurological Disorders and Stroke (NS098023).

Received: 4 November 2020; Revised: 26 December 2020; Accepted: 1 February 2021

***Annals of Clinical and Translational Neurology* 2021; 8(4): 800–810**

doi: 10.1002/acn3.51324

[a]Yuri Ahuja and Nicole Kim contributed equally as co-first authors.

[b]Tianxi Cai and Zongqi Xia contributed equally as co-senior authors.

Abstract

Objective: No relapse risk prediction tool is currently available to guide treatment selection for multiple sclerosis (MS). Leveraging electronic health record (EHR) data readily available at the point of care, we developed a clinical tool for predicting MS relapse risk. **Methods:** Using data from a clinic-based research registry and linked EHR system between 2006 and 2016, we developed models predicting relapse events from the registry in a training set ($n = 1435$) and tested the model performance in an independent validation set of MS patients ($n = 186$). This iterative process identified prior 1-year relapse history as a key predictor of future relapse but ascertaining relapse history through the labor-intensive chart review is impractical. We pursued two-stage algorithm development: (1) L_1-regularized logistic regression (LASSO) to phenotype past 1-year relapse status from contemporaneous EHR data, (2) LASSO to predict future 1-year relapse risk using imputed prior 1-year relapse status and other algorithm-selected features. **Results:** The final model, comprising age, disease duration, and imputed prior 1-year relapse history, achieved a predictive AUC and F score of 0.707 and 0.307, respectively. The performance was significantly better than the baseline model (age, sex, race/ethnicity, and disease duration) and noninferior to a model containing actual prior 1-year relapse history. The predicted risk probability declined with disease duration and age. **Conclusion:** Our novel machine-learning algorithm predicts 1-year MS relapse with accuracy comparable to other clinical prediction tools and has applicability at the point of care. This EHR-based two-stage approach of outcome prediction may have application to neurological disease beyond MS.

Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) that causes progressive neurological disability.[1] Currently available disease-modifying treatments (DMTs) for MS target neuroinflammation and delay neurodegeneration primarily by reducing inflammatory disease activity or relapse.[2,3] There is growing awareness of the long-term benefit of early initiation of DMTs.[4–6] particularly higher-efficacy DMTs in patients with a high likelihood of relapse and accelerated disability accrual consistent with aggressive MS.[7–12]

The ability to predict a patient's future relapse risk is crucial to guide the clinical decision on initiating higher-efficacy DMTs, given the trade-off of potential DMT-associated adverse events and costs. Well-established clinical predictors of future aggressive MS disease activity include older age at first neurological symptom onset, male sex, non-European descent, and importantly, frequency, and severity of prior relapse.[7,10] Additional neuroimaging and laboratory predictors of relapse include gadolinium enhancement on magnetic resonance imaging (MRI)[13] and low serum 25-OH vitamin D.[14] These factors each have modest power for predicting future relapse. While predictive models of neurological disability accrual are available,[15,16] to our knowledge, there has been no clinically deployable predictive model of future relapse that incorporates multiple predictors.

Studies that predict MS outcomes predominantly rely on research registry data. Increasing analytical capability[17,18] has enabled the use of electronic health records (EHR) data to facilitate clinical discovery by providing complementary features otherwise unavailable from traditional research registries. We previously integrated research registry data from a well-characterized, long-term, clinic-based cohort[19,20] with EHR data for developing EHR-based models of MS classification and neurological disability.[15,21] Here, we leveraged clinical and associated EHR data to develop and test a clinically deployable model for predicting 1-year relapse risk in MS patients.

Methods

Data source

We included data from January 2006 to December 2016 for 2375 participants ≥18 years of age with neurologist-confirmed MS diagnosis in the Comprehensive Longitudinal Investigation of Multiple Sclerosis at Brigham and Women's Hospital (CLIMB) cohort in the Brigham Multiple Sclerosis Center (Boston). CLIMB participants have had at least one annual clinic visit. We additionally obtained all EHR data for 5482 MS patients from the Mass General Brigham (MGB, formerly known as the Partners) HealthCare system using our published MS classification algorithm, with 4565 receiving neurological care at the Brigham MS Center.[15] The MGB IRB approved the use of research registry data and EHR data.

For the training set, we included the 1435 CLIMB participants with linked EHR data, as previously described.[15] For evaluation of model performance in a held-out validation set, we used annotated relapse events for 186 randomly selected MS patients from the EHR cohort from the same time period who received neurological care at MGB (77 in CLIMB) but were not part of the training set. We assessed for potential selection bias arising from training the model exclusively on CLIMB patients by comparing its predictive performance on the 77 CLIMB patients to the 109 non-CLIMB patients in the validation set and found no significant disparity between the subgroups. A research assistant performed the chart review according to CLIMB guidelines after extensive training and under the close supervision of an MS neurologist. Figure 1 describes the overall workflow.

Relapse data

We used relapse events, dates, and type, from the CLIMB registry (training set) and annotation (validation set). For this study, we defined a relapse event as a clinical and/or radiological relapse. Clinical relapse was defined as having new or recurrence of neurological symptoms lasting persistently for ≥24 h without fever or infection. Radiological relapse was defined as having either a new T1-enhancing lesion and/or a new or enlarging T2-FLAIR hyperintense lesion on brain, orbit, or spinal cord MRI on clinical radiology report.

EHR data

For each patient, we extracted relevant demographic and clinical information (i.e., age, sex, race/ethnicity, disease duration [years elapsed between the first MS diagnostic code and index encounter]) from the EHR data. We extracted all occurrences over time of the following codified variables: (1) diagnostic (International Classification of Disease 9th/10th edition, ICD-9/10) codes; and (2) procedural (Current Procedural Terminology, CPT) codes. Using a published classification system that consolidates multiple related ICD codes of each unique medical condition,[22] we mapped each ICD code to a single clinically informative condition represented by a "phenotype" code (PheCode).[18] To mitigate sparsity, we consolidated CPT codes according to groupings defined by the American Medical Association, with the exception of certain MRI procedures (orbit, brain, and spine) because of relevance to MS.

From free-text clinical narratives (e.g., outpatient encounters, radiology reports, discharge summaries), we extracted patient-level counts of all clinical terms mapped to concept unique identifiers (CUIs) using the Natural Language Processing (NLP)-based clinical Text Analytics and Knowledge Extraction System (cTAKES).[23] Only positive mentions of CUIs were included.

Feature selection and data preprocessing

We first derived an EHR algorithm for predicting 1-year relapse history using all available EHR features. From a list of 2726 features consisting of PheCode, CPT, and CUI occurrences within a 1-week period of a given index patient encounter, we first screened for potentially informative features by fitting marginal logistic regression models to identify features significantly associated with

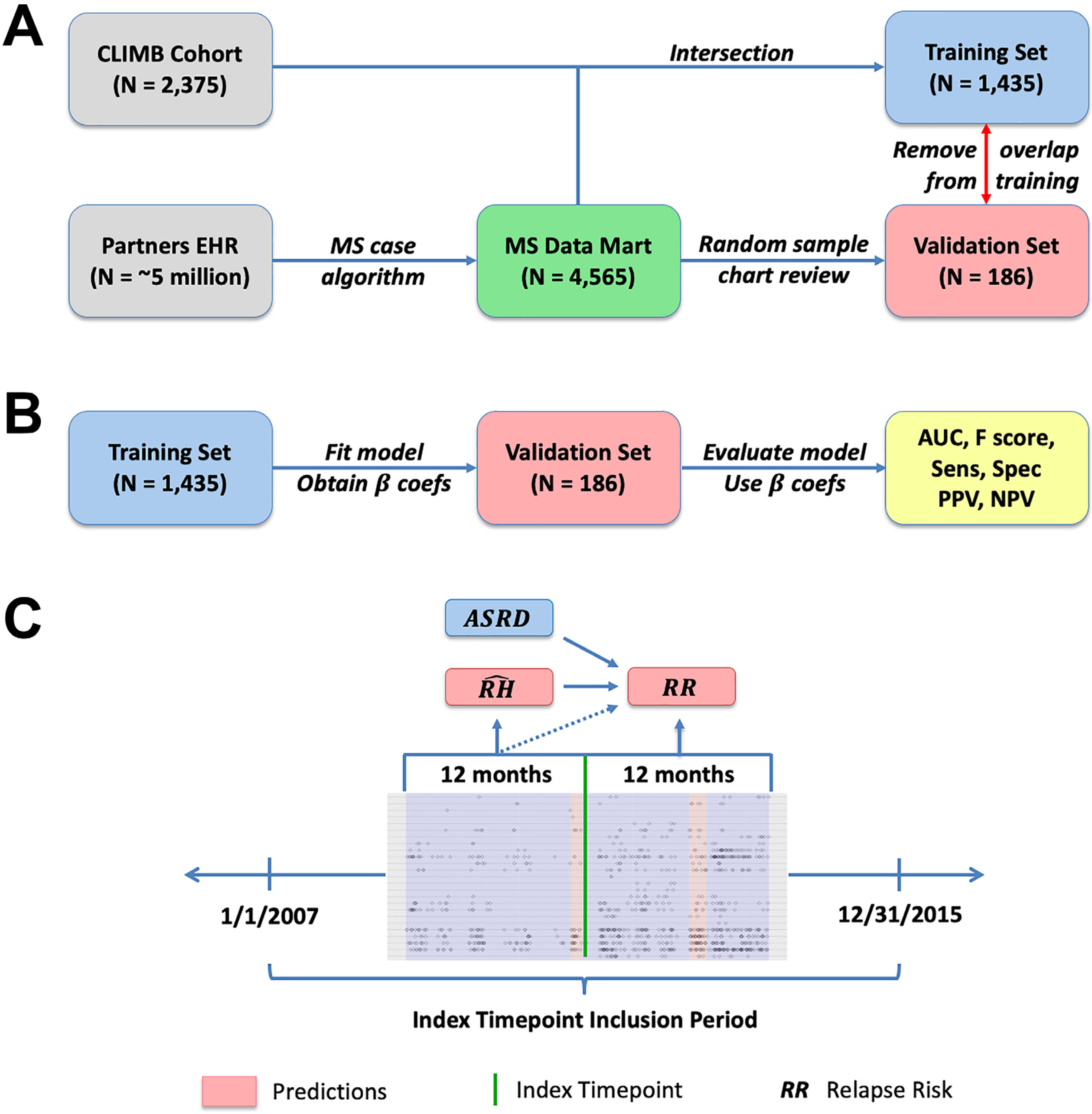

Figure 1. Study schematics. (A) data source of electronic health records and research registry data, training and validation set, (B) overall study workflow, and (C) two-stage development of phenotyping and prediction model of MS relapse risk.

relapse. We removed features with insignificant *P*-values after adjustment using the Benjamini–Hochberg procedure with a false discovery rate of 0.1.[24] This screening procedure identified a few hundred features to be included in further algorithm development. For each positively screened feature, we aggregated total counts over the prior 1-year period and log-transformed these counts. From the gold-standard CLIMB registry, we separately extracted the number of relapses each patient had in the prior 1-year period, described as 1-year relapse history (RH). We also experimented with extracting EHR data and relapse information in the prior 6-month and 2-year period. While past 1-year relapse history yielded the most accurate prediction of future 1-year relapse, predictive performance reassuringly appeared mostly insensitive to the choice of the training period length.

The main objective of the study was to predict a patient's *future* probability of relapse within one year using EHR feature counts and demographic information rather than RH, as RH is often not readily available at the point of care. To prepare algorithm development, we classified a patient encounter as a case if the patient had a relapse within 1 year after the index date, and as a control otherwise. To avoid overcounting closely occurring encounters, we randomly sampled one encounter per nonoverlapping 3-month time window for inclusion in the final preprocessed dataset. To mitigate sparsity, we eliminated features with prevalence <5% in both case and control groups. In this preprocessed dataset, each patient had multiple index timepoints over the intersection of the study inclusion period and the patient's records.

Prediction of future 1-year relapse probability

We use N, $T(i)$, and p to represent the number of patients, number of timepoints for patient i, and number of features in the preprocessed dataset, respectively. We denote the complete feature vector for patient i in time period (t-1y, t) as $X_{i,t} = (X_{i,t,1},\ldots, X_{i,t,p})'$ and $X_i = (X_{i,1},\ldots, X_{i,T(i)})'$. We let $X_i = (ASRD_i, EHR_i)$, where $ASRD_{\mathrm{i}}$ denotes age, sex, race/ethnicity, and disease duration, whereas EHR_i denotes aggregated EHR features in the prior 1-year period. Furthermore, we use $Y_{i,t}$ to indicate whether patient i has a relapse in time period (t, t + 1y), and let $Y_i = (Y_{i,1}, \ldots, Y_{i,T}(i))'$. Finally, we represent the prior 1-year relapse history of patient i as $RH_{\mathrm{i,t}}$ and let $RH_{\mathrm{i}} = (RH_{\mathrm{i},1}, \ldots, RH_{i,T(i)})'$. We use ($X_{train}$, Y_{train}, RH_{train}) and (X_{test}, Y_{test}, RH_{test}) to designate nonoverlapping training and validation sets, respectively, for model development and independent evaluation (Fig. 1B).

We predicted Y_i using a two-stage procedure (Fig. 1C). In the first stage (phenotyping for RH), we predicted *contemporaneous* relapse within the *same* 1-year period as the EHR features by fitting an L_1-regularized (LASSO) linear regression to $\{X_{train}, \log\ (RH_{train} + 1)\}$. We used log (RH + 1) instead of log(RH) such that RH = 0 would yield a log relapse count of 0 rather than negative infinity. We optimized the LASSO regularization hyperparameter λ using 10-fold cross-validation to maximize Spearman correlation with the true count RH_i. We use $\widehat{RH}_{i,t}$ to denote the LASSO-predicted *past* 1-year log relapse count for patient i at timepoint t, and let $\widehat{RH}_i = \left(\widehat{RH}_{i,1},\ldots,\widehat{RH}_{i,T(i)}\right)$. We further experimented with two alternative models for imputing RH_i: (1) LASSO logistic regression predicting $I\ (RH_i > 0)$ (i.e., at least 1 relapse) and (2) LASSO Poisson regression predicting RH_i. Poisson regression assumes that the outcome follows a Poisson rather than a normal distribution (as in standard linear regression). We selected the model with the best performance to impute $\widehat{RH}$.

In the second stage (prediction), we predicted *future* 1-year relapse by fitting a LASSO logistic regression to (A) $\left\{\left(ASRD_{train},\widehat{RH}_{train}\right),Y_{train}\right\}$, and (B) $\left\{\left(X_{train},\widehat{RH}_{train}\right),Y_{train}\right\}$. Model (A) used age, sex, race, and disease duration plus $\widehat{RH}$ to predict Y, whereas Model (B) includes the features in Model (A) and all EHR features that passed the feature selection process. Importantly, neither model used the actual prior relapse history to predict future relapse, because $\widehat{RH}$ is a function of *EHR* but not *RH*.

Model evaluation

To report model performance in the validation set, we computed AUC as well as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F score, using a time-dependent threshold set at the observed prevalence among observations within ±1 year of a patient's time since the first MS relapse. AUC, sensitivity, and specificity are agnostic to outcome prevalence (which is relatively low in this study), whereas PPV, NPV, and F score (i.e., the harmonic mean of sensitivity and PPV) depend on outcome prevalence. We compared the two-stage phenotyping-prediction model to three LASSO logistic regression models trained without relapse history (model 1–3) and two models trained with relapse history (model 4–5): (1) $ASRD_i$ alone, (2) $ASRD_i$ + *MS PheCode (335)*, (3) $ASRD_i + EHR_i$, (4) $ASRD_i + RH_i$, and (5) $ASRD_i + EHR_i + RH_i$ (Fig. 2). We obtained the standard error estimates, 95% confidence intervals, and P values for comparisons of all models to the baseline *ASRD* model nonparametrically by bootstrapping with 1000 replicates.

Data availability

Code for analysis and figure generation is available at https://tinyurl.com/MS-Relapse-Prediction. Anonymous data that support the findings of this study are available upon request to the corresponding author.

Results

Patient characteristics

MS patients in the training and validation sets were comparable, specifically with respect to the percentage of women, percentage of self-reported non-Hispanic Europeans, median age at the first MS diagnosis code, and

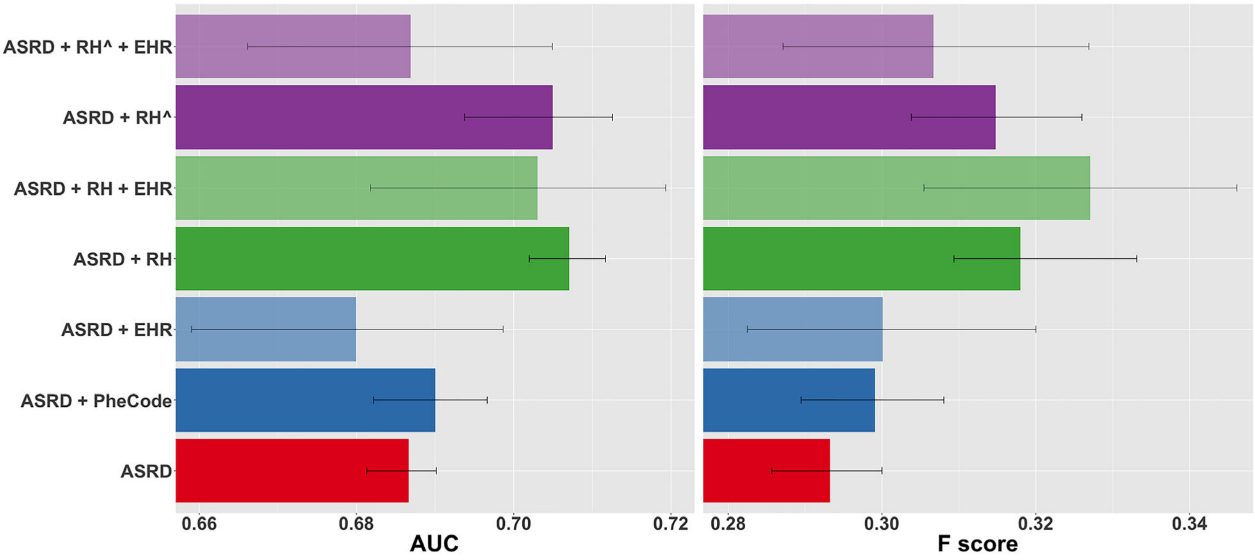

Figure 2. Performance of models in predicting the future 1-year MS relapse risk as measured by AUCs and F scores. *ASRD* (red), a baseline model comprising only basic clinical factors (age, sex, race/ethnicity, disease duration); *ASRD + PheCode* (dark blue), baseline model plus PheCode for MS; *ASRD + EHR* (light blue), baseline model plus selected EHR features that passed the feature selection process; *ASRD + RH* (dark green) and *ASRD + RH+EHR* (light green), baseline model plus actual prior 1-year relapse history without and with selected EHR features, respectively; *ASRD + RH^* (dark purple) and *ASRD + RH^+EHR* (light purple), baseline model plus two-stage phenotyping and prediction model without and with selected EHR features in the prediction stage, respectively. RH^ (equivalent to $\widehat{RH}$) denotes prior 1-year relapse history imputed from EHR data using phenotyping algorithm rather than actual relapse history (*RH*). Models were developed using the training set and evaluated on the held-out validation set. 95% confidence intervals were computed nonparametrically via bootstrap with 1000 replicates.

Table 1. Demographics of the training and validation sets.

	Training set*	Validation set*	*P*-value
Total number of patients	1435	186	NA
Sex, % Women	73.9%	74.2%	0.924
Race, % non-Hispanic European	85.9%	84.9%	0.719
Median (IQR) age at first code[1]	43.3 (15.6)	43.7 (16.0)	0.109
Median (IQR) age at first ICD code for MS	43.3 (15.5)	43.5 (16.2)	0.151
Median (IQR) disease duration, years	5.12 (2.03)	4.37 (2.82)	<0.0001
Annualized relapse rate 2006–2016, mean (SD)[2]	0.075 (0.002)	0.118 (0.009)	<0.0001

[1]The first of any ICD, CPT, or CUI code in the EHR data.

[2]Relapse type includes clinical, radiological, or both.

*The training set derives entirely from the CLIMB cohort, whereas the validation set is a random sample of MS patients from the Mass General Brigham (formerly known as the Partners) healthcare system (77 from CLIMB, none in the training set).

median age at the first occurrence of (any) ICD, CPT, or CUI code in the EHR data, whereas the disease duration was slightly shorter in the validation set (Table 1). From 2000 to 2016, the annualized relapse rate (clinical and/or radiological) was overall low, though the training set (0.074 ± 0.003) was marginally lower than the validation set (0.116 ± 0.017).

Prediction of 1-year relapse probability

The primary objective was to develop models to predict the future risk of relapse within 1 year. As measured by both AUC and F score, a model comprising basic clinical features (age, sex, race/ethnicity, and disease duration) and prior 1-year relapse history ($ASRD_i + RH_i$) performed the best in predicting future 1-year relapse risk (Fig. 2, Table 2), significantly better than the baseline $ASRD_i$ model, reflecting the important predictive value of prior relapse history. The addition of EHR features that passed the feature selection screening process to this model ($ASRD_i + RH_i + EHR_i$) diminished AUC and F score while markedly widening 95% confidence intervals (Fig. 2). Finally, the model comprising basic clinical features and selected EHR features ($ASRD_i + EHR_i$) without prior 1-year relapse history did not significantly improve AUC or F score over the baseline model ($ASRD_i$) while also widening confidence intervals.

Next, we built a phenotyping algorithm to impute past relapse history using EHR data for subsequent input into the future relapse risk prediction model, as we aimed to predict future relapse probability without using the actual

Table 2. Performance of models in predicting the future 1-year MS relapse risk.

Models[1]	AUC	P^2	*F* score	P^3	Sensitivity	Specificity	PPV	NPV
$ASRD$	0.686		0.288		0.520	0.676	0.199	0.901
$ASRD+PheCode$	0.686	0.14	0.292	0.10	0.537	0.668	0.200	0.903
$ASRD+EHR$	0.695	0.56	0.319	0.25	0.509	0.738	0.232	0.906
$ASRD+RH$	0.712	<0.01	0.339	<0.01	0.478	0.791	0.262	0.907
$ASRD+RH+EHR$	0.700	0.15	0.319	0.07	0.459	0.780	0.245	0.903
$ASRD+\widehat{RH}$	0.707	<0.01	0.307	<0.01	0.499	0.719	0.223	0.900
$ASRD+\widehat{RH}+EHR$	0.696	0.43	0.318	0.09	0.501	0.743	0.233	0.906

[1] *ASRD*, a baseline model comprising only basic clinical factors (age, sex, race/ethnicity, disease duration); *ASRD + PheCode*, baseline model plus PheCode for MS; *ASRD + EHR*, baseline model plus selected EHR features that passed the feature selection process; *ASRD + RH* and *ASRD + RH+EHR,* baseline model plus actual prior 1-year relapse history without and with selected EHR features, respectively; *ASRD + RH^* and *ASRD + RH^+EHR*, baseline model plus the two-stage phenotyping and prediction model without and with selected EHR features in the prediction stage, respectively. RH^ differs from RH in that the former denotes prior 1-year relapse history imputed from EHR data using the phenotyping algorithm, whereas the latter denotes actual prior 1-year relapse history. Models were developed using the training set and performance was evaluated on the held-out validation set. AUC and F score of all models were compared to the baseline model (ASRD).

[2] Comparison in AUC between each model and the baseline model (ASRD). *P*-values were computed nonparametrically via bootstrap with 1000 replicates.

[3] Comparison in *F* score between each model and the baseline model (ASRD). *P*-values were computed nonparametrically via bootstrap with 1000 replicates.

relapse history RH_i. With the finding that prior 1-year relapse history is an important predictor of future relapse risk, we also recognized that relapse history is often unavailable at the point of care while chart review is time consuming. In the first part of a two-stage model (phenotyping), we used selected EHR features (EHR_i) to impute contemporaneous RH_i for subsequent use in future relapse risk prediction. For this stage, LASSO linear regression achieved the highest AUC (0.790) and Spearman correlation (0.487) in the validation set (Table S1). As such, we used this model to impute $\widehat{RH}$. Among the 205 features that passed the feature selection screening process, the LASSO phenotyping algorithm selected 111 features (12 CPT codes, 60 CUIs, and 35 PheCodes) as informative of RH_i (Table S2). We found that age and disease duration were inversely associated with contemporaneous relapse. On the other hand, the CPT code for "MRI spine," PheCodes for "optic neuritis," and "other demyelinating diseases of the CNS," and CUIs for "intravenous steroid injection", "Lhermitte's sign," and "flare" were positively associated with relapse, consistent with clinical experience. When examining the Spearman correlations among the 111 selected variables (Fig. 3), we found the vast majority of features to have pairwise correlations in the range of 0–0.2, suggesting that these variables conveyed sufficiently nonredundant information.

Notably, the two-stage model comprising basic clinical features and the imputed prior 1-year relapse history based on the EHR-based phenotyping algorithm ($ASRD_i+\widehat{RH}_i$) achieved an AUC and *F* score of 0.707 and 0.307, respectively, both significantly higher than the baseline model $ASRD_i$ (AUC, $P < 0.01$; *F* score, $P < 0.01$) (Fig. 2, Table 2). Moreover, both the AUC and F score of the $ASRD_i+\widehat{RH}_i$ model were statistically noninferior to those of the model containing actual prior relapse history ($ASRD_i+RH_i$), as ascertained nonparametrically via bootstrap (AUC, $P = 0.27$; *F* score, $P = 0.38$). The ROC curve demonstrated that $ASRD_i+\widehat{RH}_i$ performed the best when setting the threshold for either high sensitivity (>~0.9) or high specificity (>~0.9), suggesting that it is best suited as either a high-sensitivity screening tool or a high-specificity prognostic algorithm (Fig. 4). The two-stage model ($ASRD_i+\widehat{RH}_i$) also exhibited markedly narrower 95% confidence intervals than $ASRD_i+EHR_i$ or $ASRD_i+RH_i+EHR_i$, suggesting that using EHR_i to impute RH_i in the phenotyping stage rather than in the final prediction model mitigated the variance-increasing effect of the high-dimensional EHR feature set (Fig. 2). The final prediction model ($ASRD_i+\widehat{RH}_i$) was driven by just three factors: age, disease duration, and prior 1-year relapse history imputed from EHR data (see coefficients in Table S3). We demonstrated sample implementations of the model as applied to one low-risk patient (who may benefit from standard-efficacy DMT or perhaps no DMT and infrequent monitoring of MS disease activity) and one high-risk patient (who may benefit from early initiation of higher-efficacy DMT and frequent monitoring of disease activity) (Fig. S2).

Calibration of relapse risk probabilities

To evaluate the utility of the two-stage model predictions as relapse *probabilities* rather than risk scores, we compared 1-year predicted relapse probability to the

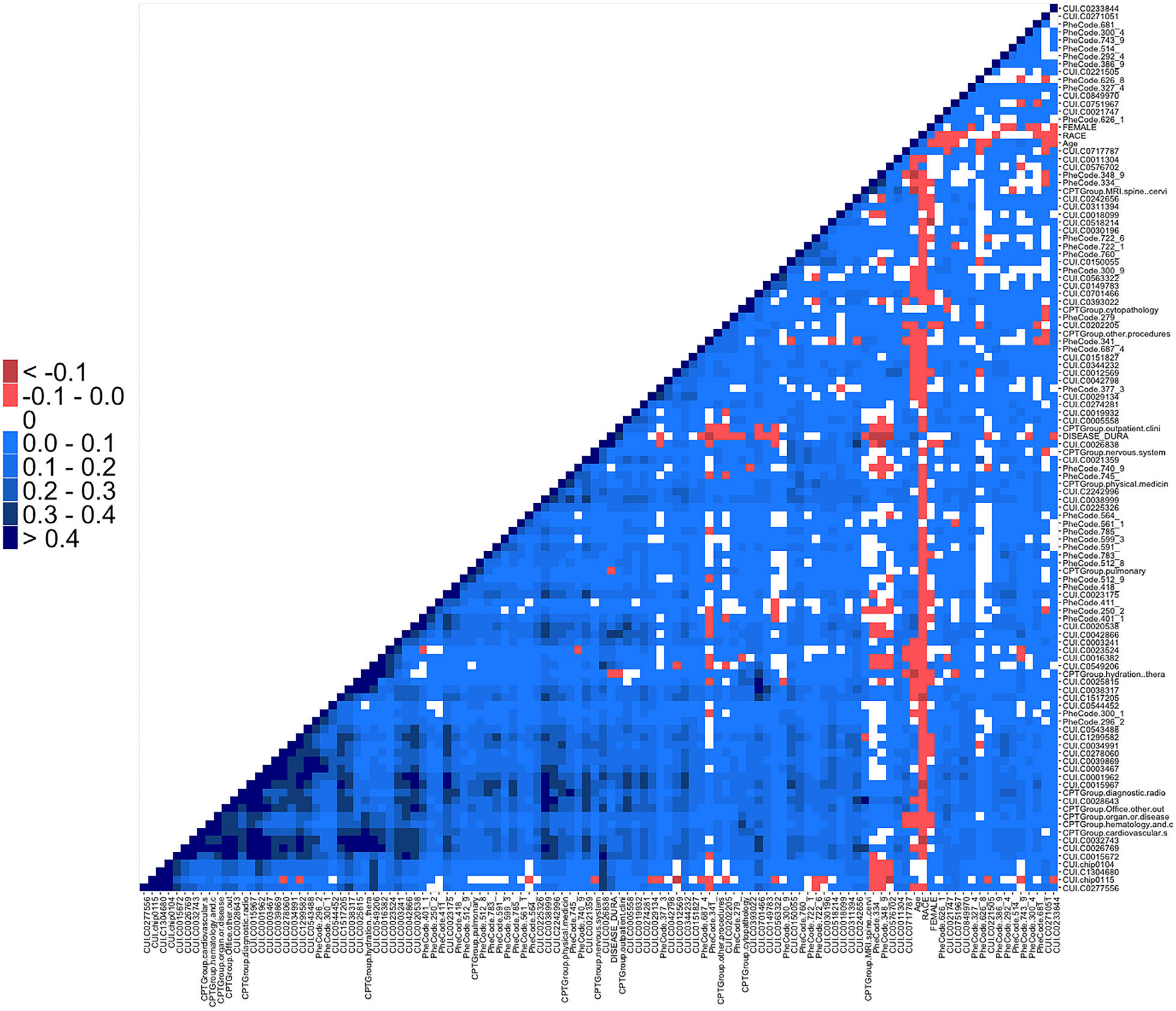

Figure 3. Heat map of pairwise correlations between prior relapse history (*RH*)-predictive features selected by LASSO in the phenotyping stage.

proportions of patients experiencing an actual relapse in the same 1-year, stratified by disease duration and patient age. We found that the actual 1-year relapse proportion declined significantly over disease duration (coefficient = −0.150, 95% CI [−0.191, −0.108], $P < 2 \times 10^{-16}$) and age (coefficient = −0.054, 95% CI [−0.065, −0.043], $P < 2 \times 10^{-16}$) (Fig. 5, Table S4), consistent with the notion that inflammatory disease activity in MS diminishes over time. In parallel, the predicted relapse probabilities also significantly declined with both disease duration (coefficient = −0.226, 95% CI [−0.242, −0.211], $P < 2 \times 10^{-16}$) and age (coefficient = −0.060, 95% CI [−0.064, −0.056], $P < 2 \times 10^{-16}$). The mean rates of decline in predicted relapse probability over age and disease duration were comparable to that of the actual relapse proportion. These results support the utility of the two-stage model as an unbiased predictor of 1-year relapse risk. By effectively leveraging EHR information to predict relapse, the two-stage model allows for a more precise, personalized prediction of risk than a predictor using age and disease duration information alone.

Supplementary analysis

We developed a two-stage model for predicting 2-year relapse risk (Tables S5 and S6, Fig. S1, Supplementary Material). While this model outperformed baseline predictors, only the AUC improvement was statistically significant. We performed two exploratory analyses to (1) demonstrate the stability of the model trained on data from 2006 to 2016 and (2) quantify the improvement of the two-stage model over baseline model in PPV and

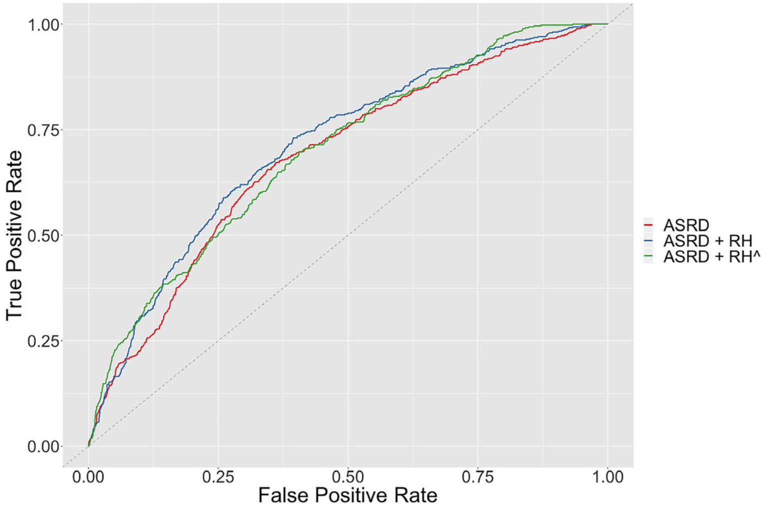

Figure 4. Receiver operating characteristic curves of models for predicting the future 1-year MS relapse probability. See Figure 2 description of *ASRD*, *ASRD + RH*, and $ASRD+\widehat{RH}$.

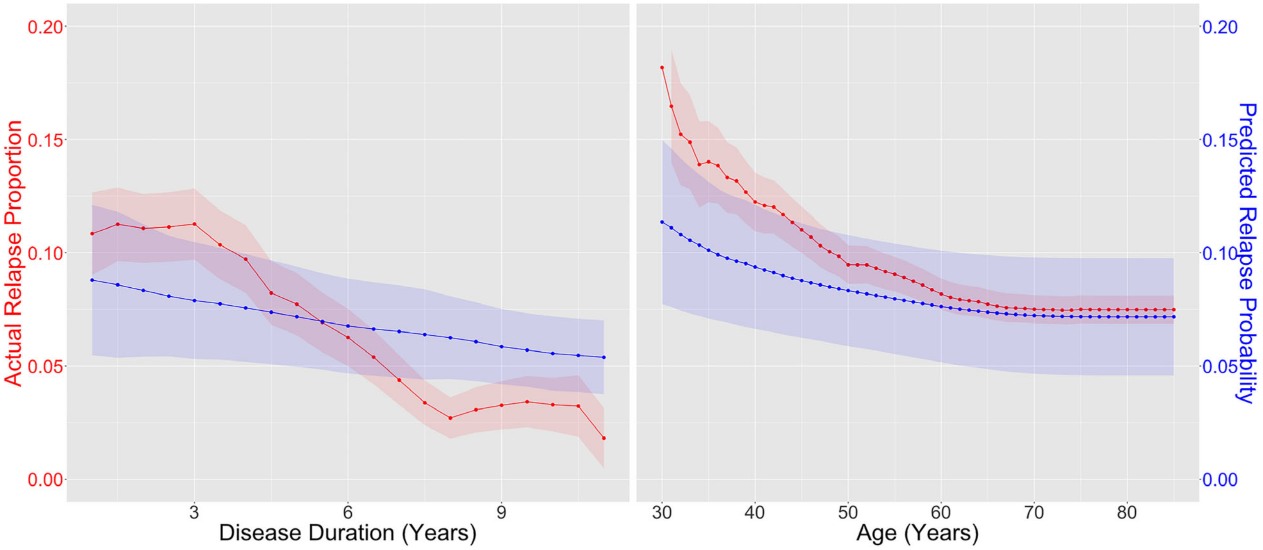

Figure 5. Relapse trend. Proportion of patients experiencing actual MS relapse (red) and mean predicted future 1-year relapse probability based on the two-stage model (blue) as a function of MS disease duration (left) and patient age (right). 95% confidence intervals for the predictive model were computed nonparametrically via bootstrap with 1000 replicates.

NPV when setting the threshold to achieve >95% specificity and >95% sensitivity as well as the number of additional high-risk and low-risk patients correctly identified per 100 tested for future 1-year relapse probability (Supplementary Material).

Discussion

The ability to accurately predict future relapse at the point of care will improve clinical decision making, particularly in selecting MS treatment. We report a novel two-stage model for predicting a patient's 1-year MS relapse risk that incorporates imputed prior relapse history based on EHR data. This model does not require knowledge of a MS patient's prior relapse frequency, a key predictor often unavailable at the point of care. Achieving clinically actionable accuracy (AUC = 0.707), this final model performed significantly better than baseline models and was noninferior to a predictive model containing actual relapse history. Furthermore, the model-predicted relapse probability declined with disease duration and patient age similar to trends seen with actual relapse proportion, suggesting that it produces a clinically meaningful estimate of a patient's relapse risk over the course of this chronic disease.

This study builds on our prior work on integrating EHR and research registry data to develop high-dimensional models for not just classifying MS diagnosis but also estimating a key measure of neurological disability in MS that is not part of routine medical records (the multiple sclerosis severity score).[15] Following the latest developments in EHR data analytics for phenotyping disease outcomes,[18] our approach leverages the rich complexity of the available EHR data by incorporating a variety of codified and narrative variables in our algorithm. The novelty of the two-stage model lies in using high-dimensional EHR data to impute the key predictor (past relapse history, which captures important aspects of individual disease profile) in the phenotyping stage rather than in the future relapse prediction stage. This method mitigates variance increase due to the high feature dimensionality of the EHR data while preserving accuracy, bypassing the practical bottleneck of the labor-intensive chart review process for ascertaining prior relapse history and improving the explicability of the final model. While the improvement over the baseline model is modest, the final prediction model achieved performance comparable to other clinical prediction algorithms. For comparison, the classic Framingham Risk Score for predicting coronary heart disease has an AUC in the 0.6–0.75 range.[25,26]

The final relapse prediction model comprised only three familiar factors (age, disease duration, and imputed number of relapses in the prior year). These predictors are consistent with prior literature.[27,28] In planning the model development, we consciously avoided including DMT history among the potential features because we plan to use the predicted relapse risk as outcomes in future analyses evaluating efficacy in reducing relapse across DMTs and because the inclusion of specific DMTs might limit its future application given the ever-growing number of DMT options. We also did not include MRI features as we originally focused on building a parsimonious model comprising clinical predictors readily available from the EHR data. We plan to incorporate MRI features in future iterations of the model.

This study faces a limitation of selection bias and potential generalizability. The relapse prediction algorithm was developed using participants from a research cohort (CLIMB) and tested on patients within the same tertiary academic hospital system (MGB). Given that routine EHR data rarely capture recorded relapse events systematically, using research registry data to train models of relapse prediction is a necessity. Additional validation in other healthcare settings is warranted. If externally validated, the relapse risk prediction model can be integrated at the point of care to systematically identify MS patients at high risk of relapse and alert clinicians in selecting the appropriate DMTs.

In summary, our novel model predicts 1-year MS relapse risk with accuracy comparable to other clinical prediction algorithms and with potential applicability at the point of care. Our EHR-based two-stage approach for MS relapse imputation and temporal relapse prediction may have application to other complex neurological outcomes apart from MS.

Conflict of Interest

The authors have declared that no conflict of interest relevant to this study exists.

References

1. Reich DS, Lucchinetti CF, Calabresi PA. Multiple sclerosis. N Engl J Med 2018;378:169–180.
2. Goodin DS, Reder AT, Bermel RA, et al. Relapses in multiple sclerosis: relationship to disability. Mult Scler Relat Disord 2016;6:10–20. https://doi.org/10.1016/j.msard.2015.09.002
3. Baecher-Allan C, Kaskow BJ, Weiner HL. Multiple sclerosis: mechanisms and immunotherapy. Neuron 2018;97:742–768. https://doi.org/10.1016/j.neuron.2018.01.021
4. Kappos L, Freedman MS, Polman CH, et al. Long-term effect of early treatment with interferon beta-1b after a first clinical event suggestive of multiple sclerosis: 5-year active treatment extension of the phase 3 BENEFIT trial. Lancet Neurol 2009;8:987–997. https://doi.org/10.1016/S1474-4422(09)70237-6
5. Chalmer TA, Baggesen LM, Nørgaard M, et al. Early versus later treatment start in multiple sclerosis: a register-based cohort study. Eur J Neurol 2018;25:1262. https://doi.org/10.1111/ene.13692
6. Brown JWL, Coles A, Horakova D, et al. Association of initial disease-modifying therapy with later conversion to secondary progressive multiple sclerosis. JAMA 2019;321:175–187. https://doi.org/10.1001/jama.2018.20588
7. Rush CA, MacLean HJ, Freedman MS. Aggressive multiple sclerosis: proposed definition and treatment algorithm. Nat Rev Neurol 2015;11:379–389. https://doi.org/10.1038/nrneurol.2015.85
8. Giovannoni G. Disease-modifying treatments for early and advanced multiple sclerosis: a new treatment paradigm. Curr Opin Neurol 2018;31:233–243. https://doi.org/10.1097/WCO.0000000000000561
9. Harding K, Williams O, Willis M, et al. Clinical outcomes of escalation vs early intensive disease-modifying therapy in patients with multiple sclerosis. JAMA Neurol

2019;76:536–541. https://doi.org/10.1001/jamaneurol.2018.4905

10. Ontaneda D, Tallantyre E, Kalincik T, et al. Early highly effective versus escalation treatment approaches in relapsing multiple sclerosis. Lancet Neurol 2019;18:973–980. https://doi.org/10.1016/S1474-4422(19)30151-6
11. Bose G, Freedman MS. Precision medicine in the multiple sclerosis clinic: selecting the right patient for the right treatment. Mult Scler J 2020;26:540–547. https://doi.org/10.1177/1352458519887324
12. He A, Merkel B, Brown JWL, et al. Timing of high-efficacy therapy for multiple sclerosis: a retrospective observational cohort study. Lancet Neurol 2020;19:307–316. https://doi.org/10.1016/S1474-4422(20)30067-3
13. Kappos L, Moeri D, Radue EW, et al. Predictive value of gadolinium-enhanced magnetic resonance imaging for relapse rate and changes in disability or impairment in multiple sclerosis: a meta-analysis. Gadolinium MRI Meta-analysis Group. Lancet 1999;353:964–969.
14. University of California, San Francisco MS-EPIC Team; Cree BAC, Gourraud P-A, Oksenberg JR, et al. Long-term evolution of multiple sclerosis disability in the treatment era. Ann Neurol 2016;80:499–510. https://doi.org/10.1002/ana.24747
15. Xia Z, Secor E, Chibnik LB, et al. Modeling disease severity in multiple sclerosis using electronic health records. PLoS One 2013;8:e78927. https://doi.org/10.1371/journal.pone.0078927
16. Malpas CB, Manouchehrinia A, Sharmin S, et al. Early clinical markers of aggressive multiple sclerosis. Brain 2020;143:1400–1413. https://doi.org/10.1093/brain/awaa081
17. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885. https://doi.org/10.1136/bmj.h1885
18. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc. 2019;14:3426–3444. https://doi.org/10.1038/s41596-019-0227-6
19. Gauthier SA, Glanz BI, Mandel M, Weiner HL. A model for the comprehensive investigation of a chronic autoimmune disease: the multiple sclerosis CLIMB study. Autoimmun Rev 2006;5:532–536. https://doi.org/10.1016/j.autrev.2006.02.012
20. Rotstein DL, Healy BC, Malik MT, et al. Evaluation of no evidence of disease activity in a 7-year longitudinal multiple sclerosis cohort. JAMA Neurol 2015;72:152–158. https://doi.org/10.1001/jamaneurol.2014.3537
21. Zhang T, Goodman M, Zhu F, et al. Phenome-wide examination of comorbidity burden and multiple sclerosis disease severity. Neurology: Neuroimmunol Neuroinflamm 2020;7:e864. https://doi.org/10.1212/NXI.0000000000000864
22. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics 2014;30:2375–2376. https://doi.org/10.1093/bioinformatics/btu197
23. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507–513. https://doi.org/10.1136/jamia.2009.001560
24. Yoav B, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol) 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
25. Wilson PW, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837–1847.
26. Orford JL, Sesso HD, Stedman M, et al. A comparison of the Framingham and European Society of Cardiology coronary heart disease risk prediction models in the normative aging study. Am Heart J 2002;144:95–100. https://doi.org/10.1067/mhj.2002.123317
27. Fromont A, Debouverie M, Le Teuff G, et al. Clinical parameters to predict response to interferon in relapsing multiple sclerosis. Neuroepidemiology 2008;31:150–156. https://doi.org/10.1159/000151524
28. Kalincik T, Manouchehrinia A, Sobisek L, et al. Towards personalized therapy for multiple sclerosis: prediction of individual treatment response. Brain 2017;140:2426–2443. https://doi.org/10.1093/brain/awx185

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

File S1. Supplementary material on the development and performance of a two-stage model for predicting future 2-year relapse probability, exploratory analysis of the stability of the two-stage future 1-year relapse probability prediction model from 2006 to 2016, and exploratory analysis of the improvement of the two-stage model future 1-year relapse probability prediction model over the baseline model.
Table S1. Performance of models for the phenotyping or prediction of contemporaneous 1-year MS relapse history.
Table S2. LASSO linear regression coefficients of the phenotyping algorithm for imputing contemporaneous 1-year relapse history in the two-stage model.

Table S3. LASSO logistic regression coefficients of the prediction algorithm for future 1-year relapse probability in the two-stage model.

Table S4. Logistic regression coefficients and Spearman correlations of actual relapse proportions and predicted relapse probabilities predicted by the two-stage ASRD $+\widehat{RH}$ model over disease duration and age.

Table S5. Performance of models in predicting the future 2-year MS relapse risk.

Table S6. Performance of models for the phenotyping stage of imputing contemporaneous 2-year MS relapse history.

Figure S1. Performance of models in predicting the future 2-year MS relapse risk as measured by AUCs and F scores.

Figure S2. Sample implementation of the two-stage model of relapse prediction in two representative patients.