Ensemble genetic machine learning identifies Multiple Sclerosis genetic loci associated with future worsening of disability

Valery Fuh-Ngwa
University of Tasmania
Yuan Zhou
University of Tasmania
Phillip E. Melton
University of Tasmania
Ingrid van der Mei
University of Tasmania
Jac C. Charlesworth
University of Tasmania
Xin Lin
University of Tasmania
Amin Zarghami
University of Tasmania
Simon A. Broadley
Griffith University
Anne-Louise Ponsonby
University of Melbourne Murdoch Children's Research Institute, Royal Children's Hospital
Steve Simpson-Yap
The University of Melbourne
Jeannette Lechner-Scott
University of Newcastle
Bruce V. Taylor (bruce.taylor@utas.edu.au)
University of Tasmania

Article

Keywords: Multiple Sclerosis, Disability Progression, Genetic Variations, Ensemble Machine Learning, Genetic Decision Rules.

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1609681/v1

Abstract

Limited studies have been conducted to identify and validate multiple sclerosis (MS) genetic loci associated with disability progression. We aimed to identify MS genetic loci associated with worsening of disability over time, and to develop and validate ensemble genetic learning model(s) to identify people with MS (PwMS) at risk of future worsening. We examined associations of *208* previously established MS genetic loci with the risk of worsening of disability; we learned ensemble genetic decision rules and validated the predictions in an external dataset. We found 7 genetic loci (*rs7731626: HR = 0.92, P = 2.4x10^{-5}*; *rs12211604: HR = 1.16, P = 3.2x10^{-7}*; *rs55858457: HR = 0.93, P = 3.7x10^{-7}*; *rs10271373: HR = 0.90, P = 1.1x10^{-7}*; *rs11256593: HR = 1.13, P = 5.1x10^{-57}*; *rs12588969: HR = 1.10, P = 2.1x10^{-10}*; *rs1465697: HR = 1.09, P = 1.7x10^{-128}*) associated with worsening of disability; most of which are located near or tagged to *13* genomic regions enriched in peptide hormones and steroids biosynthesis pathways by positional and *eQTL* mapping. The derived ensembles provided a set of genetic decision rules that can be translated to provide additional prognostic values to existing clinical predictions, with the additional benefit of incorporating relevant genetic information into clinical decision making for PwMS. The present study extends our knowledge of MS progression genetics and provides the basis of future studies regarding the functional significance of the identified loci.

Introduction

Multiple Sclerosis (MS) is a chronic neurodegenerative disease typified by the accumulation of disability at varying rates[1]. MS occurs in people who have an underlying genetic susceptibility and are exposed to viral and environmental risk factors.[2] Genetic determinants of the progression of disability in MS remain elusive. Although the International MS Genetic Consortium (IMSGC)[3,4] have identified $\sim$ *233* genetic loci to be associated with MS risk, limited studies have been conducted to identify those that predict future worsening of disability.[5–9]

Machine learning models have recently been applied in studies of MS disability progression, including classical random forest (RF) and gradient boosting machines (GBM).[5,10–19] Despite their continued use in predicting MS progression outcomes, past and recent studies[15,16] (not related to MS) have shown that these models have (1) *limited clinical utility*: as they rely strictly on a discrete-time evolution of disease processes; meanwhile in MS, disability progression is characterised by a continuous-time evolution of expanded disability status scores (EDSS)[1,9]; (2) *weak predictive power*: as they do not account for correlated outcomes[20] e.g. correlation due to the sporadic time series of EDSS[17,19]; (3) *lack interpretability*: as it is difficult to understand how such models make prediction decisions.[21] Based on lessons learned from precision medicine,[22] RF and GBM are prone to overfitting and selection bias as their internal variable-splitting mechanisms often generates variables with too many possible splits/choices. They also rely on the property of *normality, independent*, and *identically* distributed outcomes, which are frequently violated in real-world applications. In addition to RF and GBM classifiers, support vector machines, neural network, and deep learning algorithms have similar drawbacks.[20,22]

Recently, Ngufor *et al.*[20] developed a mixed-effect machine learning (MEML) platform that combined the properties of RF and GBM with generalised mixed-effects regression trees (GMERT) to predict changes in glycaemic control for patients with Type 2 diabetes. Compared to the RF and GBM, their mixed-effect counterparts called MErf and MEgbm, respectively, can learn from complex tree ensembles to produce simple, readable, and interpretable risk models to assist in clinical predictions.[14,20] By definition, a MErf or MEgbm is a combination of GMERT with RF or GBM, respectively. These models are called MEML models; and have been shown to have better predictive accuracy and sensitivity compared to classical RF and GBM.[14,20] *MEML* methods may ameliorate the aforementioned limitations of classical algorithms used in predicting longitudinal changes in EDSS mediated by MS genetic markers such as single nucleotide polymorphisms (SNPs).

Using well-established MS genetic loci (risk SNPs) from IMSGC[3,4], we aimed to identify MS genetic loci associated with worsening of disability over time; and to develop and validate simple, learnable, and interpretable ensemble genetic learning model(s) to identify people with MS (PwMS) at risk of future worsening. To this end, we hypothesised that disability accumulation on the EDSS scale follows a first-order Markov process in which future disability is predicated on the prior disability history, and genetic predisposition.

Materials And Methods

Data, study cohort, and Inclusion Criteria

Using data pooled from the multi-centre (Brisbane, Newcastle, Geelong and Western Victoria, and Tasmania) Australian Longitudinal Prospective Cohort Study (AUSLONG) of MS, we analysed *279* prospectively assessed first demyelination event (FDE) participants enrolled between 2003 and 2006.[23] The AUSLONG study has ethical approval from the Tasmanian Health and Medical Research Ethics Committee (ref: H0010499, 01/-5/2009); the Queensland Institute of Medical Research Human Research Ethics Committee (ref: P1252, 22/05/2009); the Royal Brisbane and Women's Hospital Human Research Ethics Committee (ref: HREC/09/QRBW/299, 19/10/2009); the Hunter New England Human Research Ethics Committee (ref: 09/04/15/5.04, HREC/09/HNE/139, SSA/09/HNE/140, 10/08/2009); and the Barwon Health Human Research Ethics Committee (ref: BH 09/24, BH 03/46, 04/08/2009). All experiments (blood collection, genotyping, and clinical examinations) were conducted in accordance with the guidelines of each committee at the participating centres. Written informed consent was obtained from all subjects and/or their legal guardian(s) in accordance with the Declaration of Helsinki.[24] EDSS scores were acquired prospectively at intervals up to *15* years post FDE by trained and certified neurologists, and a validated telephone EDSS was obtained at yearly computer-assisted telephone interviews from *2–3* years post FDE. Initial data extraction (*n = 279* cases) was done using the revised 2017 McDonalds criteria[25] in which cases were defined as either clinically isolated syndrome (CIS), relapsing-onset MS (ROMS), secondary progressive MS (SPMS), or progressive-onset MS (POMS). The selection criteria for the final cohort were done as illustrated in Fig. 1.

Genotyping, imputation, and quality control

All participant samples were genotyped using the Illumina Infinium Global Screening Array-24 v2.0 BeadChip. Genotypes were called using Illumina GenomeStudio software. To maximise genetic coverage, the dataset was imputed using the algorithm implemented in IMPUTE version 4[26] using 1000 Genomes phase 3[4] as the reference genotype. Strict quality control was applied using established protocols[27]. SNPs previously identified as related to MS risk by IMSGC[3,4] were considered in the association analysis.

Imputation of missing EDSS measures

Imputation of missing EDSS (n = 471 of 3065) was based on a Bayesian approach using the *JointAI* R-package[28]. Conditional on the observed SNPs genotypes, the EDSS scores were considered missing at random. The imputation model is a mixed-effect proportional odds model[1] defined on 8 disability states (**1** = *EDSS [0-1.5]*, **2** = *EDSS [2-2.5]*, **3** = *EDSS [3-3.5]*, **4** = *EDSS [4-4.5]*, **5** = *EDSS [5-5.5]*, **6** = *EDSS [6-6.5]*, **7** = *EDSS [7-7.5]*, and **8** = *EDSS [8-9.5]*). Based on the results from ours[9] and previous studies[29–33], clinical and environmental factors e.g. sex, age at disease onset, body mass index (BMI), titre of Epstein-barr Nuclear Antigen IgG (EBNA), smoking status, hospital anxiety depression scores (HADS), and previous EDSS scores (EDSSPREV), were used as covariates in the imputation model following their importance in predicting worsening of disability.

Outcome measures, and data structure

Based on the study design, a first-order Markov's assumption for continuous-time EDSS transitions was considered,[1,34–36] and defined as: "*the current EDSS state depends on the previous state (EDSSPREV), and all covariate histories*". In other words, using 8 categories (listed above) of the newly imputed EDSS states, we considered a continuous-time evolution for each disability state, wherein the state at the previous observation is retained until the current visit. Note that an observation may also represent a transition to a different state before arriving at the current state, or a repeated observation of the same underlying state at the end of follow-up.

Using these assumptions, we transformed the data and defined our clinical endpoint to capture continuous-time changes in disability states as: *y = 1* denoting a "*worsening*" event *(progressive transitions)* made by an individual from study entry, and *y = 0* denoting "*improved*" events, respectively. Since individuals entered the study at different times, we defined the *time-to-worsening* of disability *t*, as the continuous time elapsed since MS diagnosis until the current observation. This was achieved using the "*msm2Surv*" function in *mstate* R-package.[37] At study entry, *t* = 0 was defined for all cases to enable comparison of the baseline hazards. All stable-state transitions (*no change in EDSS*) were excluded as these were considered non-informative censoring events,

and could lead to the *likelihood drainage*, and potentially alter the results. Thus, with regards to censoring, only informative ("*improved*") events were censored.

Statistical analysis

To identify risk SNPs that predicted the *time-to-worsening*, and/or associated with future *worsening* events over time, a three-stage process was employed.

Stage 1: Variable selection and risk estimation. We randomly split the genotype data into *75%* training (*N* = 202), and *25%* test cohorts (*N* = 67) as depicted in Fig. 2. Utilising the training cohort, we identified SNPs (including interactions with EDSSPREV) that had independent associations with the hazards of worsening from three widely used penalised Cox models, namely: least absolute shrinkage and a selection operator (*LASSO*), elastic net (*EN*), and non-negative garrotte combined with sure independent screening (*NNG-SIS*), using *10*-fold cross-validation (CV). *LASSO* and *EN* were fitted using the Goeman's *penalised* R-package[38], and *NNG-SIS* using customised functions[39]. SNPs having non-zero effects were retained. Unbiased effect sizes for the selected SNPs were then estimated by fitting continuous-time mixed-effect multivariable Cox models on the training cohort,[40,41] and using $p \leq 0.05$ *with* backward elimination,[9,40] candidate SNPs were retained in the final model. Next, a functional mapping was conducted using the FUMA software[42].

Stage 2: Constructing genetic risk ensembles. To obtain learnable predictions, we trained widely used RF and GBM models using the candidate SNPs selected from stage 1 and compared their performance with MErf and MEgbm using time-dynamic ROC analysis. This was achieved using internal and external *5*-fold cross-validation on the training cohort. That is, we split the entire training cohort into dynamic lagged training and internal validation sets, such that predictors in the current visit were used to predict outcomes in the next visit[20]. To determine whether a SNP was directly associated with a worsening event, we fitted mixed-effects logistic decision trees with the top candidates, and extracted decision rules using the *"inTrees"* (interpretable trees) algorithm in the *MEml* R-package.[20] To adjust for differences in individual progression rates, and progression rates due to MS disease course phenotypes, we included individual- and phenotype-level random effects, respectively.

Stage 3: Validation of the ensembles and their prediction rules. To evaluate the ensembles and the generated decision rules, we assessed externally the performance of the ensembles on the test cohort. Time-dynamic ROC (receiver operating characteristic curves) analysis was used to assess how well each model predicted future worsening events. The importance of the SNPs and their genetic decision rules in predicting future worsening of disability was estimated and evaluated on the training and test cohorts, respectively. We prioritised SNPs based on average Gini impurity and relative influence; and by their deleteriousness in the human genome using combined annotation dependent depletion (CADD) scores.[43]

Data availability

The raw data (genotypes and phenotypes) that enabled the findings of this study were made available by the AUSLONG/AUSIMMUNE investigators research group (https://www.msaustralia.org.au/ausimmune/). They were accessed under approval number H0010499 and are not publicly available. Please contact the corresponding author to arrange data access. GWAS summary statistics for all the genotypes is available on the IMSGC website and has been uploaded to dbGaP (https://www.ncbi.nlm.nih.gov/gap/ddb/).

Results

We analysed at total of *269* FDE cases with *2786* EDSS transitions, with subsequent diagnosis as ROMS (n = 149), POMS (n = 12); SPMS (74), while 34 remained as CIS by the 10th year review. Of these, 76.8% (n = 205) were females, and the mean age at study entry was 37.5 years (SD = 9.9 years). Of the initial 279 cases (Fig. 1), 10 cases (seen once) were excluded from the analysis.

Observed versus imputed EDSS transitions

The distribution of the observed transition percentages before and after imputation are shown in Table 1. Previous disability states (EDSSPREV) were key determinants of future states, and thus satisfies the first-order Markov process described above. (Web Appendix A1). Overall, all variables in the imputation model (mentioned above) were imputed with high accuracy judging from

Gelman-Rubin's diagnostic criteria (GR-crit ≤ 1.1, Web Appendix A1), the density plots and the mixing rates of the Markov chains (Web Appendix A2).

Table 1
EDSS transition percentages (%) before(after) imputation

Number of MS subjects = 269, number of transitions before(after) imputation = 2029(2786)									
To EDSS									
From EDSS	**0-1.5**	**2-2.5**	**3-3.5**	**4-4.5**	**5-5.5**	**6-6.5**	**7-7.5**	**8-9.5**	**Total**
0-1.5	51.7(39.8)	11.4(11.7)	14.9 (17.5)	18.6(24.9)	2.7(4.9)	0.6(1.2)	0(0)	0(0)	**516(691)**
2-2.5	13.3(13.6)	40.5(32.3)	15.0(12.3)	25.1(26.8)	4.9(9.0)	1.4(5.9)	0(0)	0(0)	**346(455)**
3-3.5	18.8(21.4)	12.8(13.8)	44.6(37.2)	17.0(18.3)	5.4(5.4)	1.2(3.6)	0(0)	0.3(0.2)	**336(443)**
4-4.5	14.4(20.8)	15.1(14.7)	12.4(13.0)	46.0(39.5)	9.6(9.3)	2.4(2.6)	0(0)	0(0)	**450(645)**
5-5.5	7.7(14.9)	8.8(9.7)	6.6(8.9)	14.4(16.0)	48.9(37.2)	12.6(13.0)	0(0)	0.5(0.4)	**182(269)**
6-6.5	0.5(1.5)	0.5(5.0)	0(4.2)	5.9(9.7)	8.1(8.5)	81.7(66.0)	1.6(3.1)	1.6(1.9)	**186(259)**
7-7.5	0(0)	0 (0)	0(0)	0(0)	0(16.7)	80.0(58.3)	20.0(8.3)	0(16.7)	**5(12)**
8-9.5	0(0)	0(0)	0(0)	0(8.3)	0(8.3)	0(8.3)	25.0(25.0)	75.0(50.0)	**8(12)**
The disability states were recoded as: 1 *= [0-1.5]*, 2 *= [2-2.5]*, 3 *= [3-3.5]*, 4 *= [4-4.5]*, 5 *= [5-5.5]*, 6 *= [6-6.5]*, 7 *= [7-7.5]*, 8 *= [8-9.5]*									

Identification and annotation of candidate genetic effects

After QC, *208* of the *233* list of MS risk loci from IMSGC[3,4] (including *rs3129889* that tags *HLA-DRB1*1501* genotype) entered stage 1. The number of variables that overlapped between the screening methods is presented in Fig. 3(a). *LASSO* and *EN* produced very similar results within *10*-fold CV on the training cohort, whereas *NNG-SIS* identified *86* SNPs associations. The final model (Table 2) included *28* variants ($p \leq 0.05$) that were in proximity to *33* unique genes.

Table 2
A genetic ensemble model for predicting disability progression in multiple sclerosis.

SNP	*CHR*	*POS*	*Alleles*	*Region*	*Nearest Gene*	*P-value*	*HR*	*β*	*SE*	*avIMP*	*CADD*
*rs*61863928 γ_7	10	64449549	G/T	Exon	*ADO*	1.2e-04	0.94	-0.07	0.02	1.00	15.36
*rs*12722559	10	6070273	C/A	Up	*IL2RA*	6.1e-04	0.90	-0.11	0.03	0.94	6.50
*rs*4808760	19	18301979	G/C	Up	*IL12RB1*	3.8e-02	0.93	-0.07	0.03	0.86	8.82
*rs*9277626	6	33081823	G/A	Exon	*DPB2*	2.2e-03	0.94	-0.07	0.02	0.85	9.03
*rs*12434551	14	69253364	A/T	Exon	*ZFP36L1*	4.0e-02	1.07	0.07	0.03	0.85	2.39
*rs*7260482	19	45143942	A/C	Exon	*PVR*	1.3e-01	0.91	-0.10	0.06	0.85	1.15
*rs*12588969	14	103230758	C/G	Exon	*RCOR1*	2.1e-10	1.10	0.09	0.01	0.84	15.09
*rs*6032662	20	44734310	C/T	Exon	*SLC12A5*	9.3e-02	1.11	0.10	0.06	0.84	5.67
*rs*11256593	10	6117322	T/C	Up	*IL15RA*	5.1e-57	1.13	0.12	0.01	0.84	1.28
*rs*802730	6	128280104	T/C	Exon	*THEMIS*	1.1e-02	1.06	0.06	0.02	0.83	11.86
*rs*962052	2	151644203	C/T	Exon	*RBM43*	9.7e-02	1.07	0.07	0.04	0.83	3.37
*rs*1465697	19	49837246	C/T	Up	*DKKL1*	1.7e-128	1.09	0.09	0.00	0.83	2.10
*rs*2590438	3	187565968	T/G	Exon	*BCL6*	3.6e-03	1.09	0.09	0.03	0.83	1.45
*rs*1801133	1	11856378	A/G	Missense	*MTHFR*	1.7e-02	0.94	-0.06	0.03	0.80	25.6
*rs*11852059	14	52306091	A/C	Up	*FRMD6*	8.8e-02	0.93	-0.07	0.04	0.80	5.06
*rs*531612	11	65705432	C/T	Exon	*EHBP1L1*	3.0e-02	1.08	0.08	0.04	0.80	0.13
*rs*12925972	16	79111297	C/T	Intron	*DYNLRB2*	6.6e-04	1.09	0.09	0.03	0.79	8.26
*rs*1177228	2	61242410	G/A	Up	*PUS10*	4.1e-03	1.09	0.09	0.03	0.79	0.08
*rs*2286974	16	11114512	G/A	Exon	*CLEC16A*	8.5e-02	1.11	0.11	0.06	0.77	0.43
*rs*2705616	4	87862396	G/C	Intron	*AFF1*	2.3e-04	0.89	-0.11	0.03	0.76	3.36
*rs*58166386	19	16559421	G/A	Intron	*EPS15R*	1.5e-04	1.08	0.08	0.02	0.75	0.14
*rs*10271373	7	138729795	C/A	UTR-3	*ZC3HAV1*	1.1e-07	0.90	-0.10	0.02	0.74	10.90
*rs*72989863	4	164493807	G/A	Intron	*MARCH1*	4.1e-03	1.07	0.07	0.02	0.73	0.24
*rs*55858457 γ_5	7	2443302	G/T	Up	*CHST12*	3.7e-07	0.93	-0.07	0.01	0.71	2.06
*rs*12211604	6	7100029	A/G	Up	*RREB1*	3.2e-07	1.16	0.15	0.03	0.63	0.05
*rs*6533052 γ_3	4	103911781	A/G	Up	*SLC9B1*	4.4e-03	0.95	-0.05	0.02	0.62	2.12
*rs*7731626 γ_2	5	55444683	G/A	Intron	*ANKRD55*	2.4e-05	0.92	-0.08	0.02	0.50	1.37
*rs*6589939 γ_4	11	122518525	A/G	Intron	*UBASH3B*	3.0e-02	0.93	-0.07	0.03	0.41	1.39

Transition-specific SNPs have superscripts $\gamma_{(.)}$ indicating their interaction with previous EDSS states (edssprev). The previous EDSS states are define by the parameters: γ_1: edssprev = 1, γ_2: edssprev = 2; γ_3: edssprev = 3 γ_4: edssprev = 4; γ_5: edssprev = 5, γ_6: edssprev = 6, γ_7: edssprev = 7, γ_8: edssprev = 8.

The volcano plot (Fig. 3(b)) reveals *12* SNPs having *p-values* below the family-wise threshold ($p = 0.002$). Most of these had minor allele frequencies > 10%. We observed $\simeq 44$%, and $\simeq 56$% differences in progression rates (intra-class correlations) between MS participants due to repeated EDSS measurements, and MS disease course phenotypes, respectively. Functional gene-enrichment using FUMA software[42] revealed *7* lead SNPs (*rs7731626*, *rs12211604*, *rs55858457*, *rs10271373*, *rs11256593*, *rs12588969*, *rs1465697*), some of which were located near, or tagged to one of *13* genes enriched in peptide hormones and steroids biosynthesis (Web Appendix B).

Interpretation of the validated ensembles

MErf and MEgbm had the highest accuracies in the training and testing cohorts (Web Appendix C1) and are best suited to describing subject characteristics that influence disability progression over time. Particularly, as the number of repeated observations increased with time, we observed better performance using MErf and MEgbm (Web Appendix C2), whereas the performance of RF and GBM deteriorated. Although all ensembles used identical marker sets, the increased performance observed with MErf and MEgbm indicated that these methods take advantage of the increasing sample size and correlation induced by multiple observations within a subject to yield more robust, models. Overall, the ensemble-derived predicted disability worsening outcomes correlated well with the observed outcomes in both the training ($r = 0.90$, $p = 2.2x10^{-16}$) and testing ($r = 0.86$, $p = 2.2x10^{-16}$) cohorts, respectively (Fig. 4). Additionally, we found consistent results conditional on MS disease course phenotype (Fig. 4). Figure 5 shows the relative importance scores (scaled *0* to *1*) in predicting disability, with significant changes observed over time. Notably, none of the top 7 SNPS has been shown to have a functional role in MS disability accrual.

Extraction and interpretation of genetic decision rules

MEgbm and MErf ensembles produced identical sets of decision rules constructed using the *28* SNP candidates. Web appendix C3 shows a decision tree of the top genetic decision rules extracted from both methods (only1st, 2nd, 3rd, and 4th are shown). We selected all rules of length between 2 and 5, with frequency < 0.01, and error < 0.25 in predicting future disability (shown in the leafs). The importance of these rules in the training and testing cohorts are shown at the end of the leafs at each visit. These rules indicate how frequent the individual ensembles trees combine a set of influential genetic variants among the *28* candidates to make prediction decisions regarding future disability status for a person living with MS. For instance, during the 1st clinical visit, the MErf ensemble uses *2* (*rs9277626* and *rs7731626*) of the *28* SNPs candidates to correctly identified participants of different MS subtypes in the training cohort prone to future worsening of disability (score = *100%*), given that their previous score was EDSS 2 or 2.5. The estimated time to progress from this state (EDSS > 2 or 2.5) was 346.8days. This rule predicted disability in 96% of MS participants in the testing cohort.

Discussion

In this study, we identified 28 significant MS genetic loci associated with risk of worsening of disability over time. Using these loci, we developed and validated simple, learnable, and robust ensemble genetic machine learning model(s) to predict disability progression, and to identify PwMS prone to future disability accumulation. However, there is little current knowledge of the functional implications of these common SNPs.

Different estimates of the variance in disability progression explained by MS genetic loci have been reported in prior studies.[5,6,8,44–47] For instance, using *125* early MS cases with *5* years of follow-up from our cohort, Pan *et al.*,[6] constructed a genetic risk score from *7* of *116* MS SNPs[3,4] to explain *32.7%* of the variance in annualised EDSS, but did not validate their findings externally; whereas Jackson *et al.*,[5] developed a RF-based genetic model on MS disease severity scores (MSSS) which included *19* of ~ *200* autosomal SNPs[3,4] to explain *21*% of the variability in MSSS, with just *4*% chance of validating their results externally. However, if we assume that the underlying assumptions made about the disability process were correct; then a common drawback to these studies is not the variability explained, but rather the utility and reliability/robustness of the identified associations, and the derived predictions in clinical practice.

In our study, we made a considered and clinically plausible Markov's assumption (i.e., that future disability is predicated on the prior disability history) to study the disability progression process in MS and employed robust MEML ensembles to predict future

disability outcomes. Of the *28* genetic loci from the IMSGC list of known MS risk loci, 7 were non-functional SNPs, and were identified as having the greatest effect. However as with MS risk, it is very difficult to provide actual biological mechanisms for the identified SNPs associations, other than just non-specific genetic markers of disability progression.

Instead of relying on the complex predictions generated by the ensembles to make prediction decisions, here we presented simple and transparent relational rules sets (see Web Appendix C3) that could be translated to aid existing clinical predictions,[10,15,17,19] or clinical research studies via a web application delivering equal prediction accuracy as the original ensemble. Clinicians could use these rules (provided genotyping was available) alongside recent clinical predictions,[10,15,17,19] and identify MS cases that are at greater risk of disability accrual in the short and medium term, and institute more aggressive MS therapies where indicated.[9]

The high intra-class correlations between the observed and ensemble-derived predicted probabilities of worsening, conditional on the SNP associations revealed a good model fit. The obtained *p-values* for these correlations (all $p \leq 5.2x10^{-10}$) were far smaller than recently reported,[5,47] suggesting a near 100% chance of replicating our results in an external MS population.

The strengths of this study lies in the assumptions we made regarding the underlying disability process in MS (defined above), and the use of novel machine learning platforms capable of analysing the longitudinal evolution of EDSS scores. By analysing the continuous-time evolution of EDSS transitions, the genetic variance in disability progression was significantly increased compared to other studies.[5,6,8,46] However, we recognise limitations in our study. For instance, our genetic ensemble lacks genome-wide coverage, and epistatic interactions amongst the MS genetic loci used. A genome-wide analysis to further identified novel SNPs associations which are not MS related, could be a fruitful area of future research. Similarly, we lack an external validation cohort (an external MS population) that matches our prospective, data dense AUSLONG cohort with genotyping available. Thirdly, as emphasised the genetic variants utilised here do not have any described biological effect making it difficult to elucidate the mechanisms underlying MS progression from these data.

In conclusion, our study provides a simple, our study provides a simple, learnable, and interpretable and robust ensemble genetic machine learning models that aggregates association evidence from 28 candidate MS genetic markers to predict future worsening of disability in PwMS. Our ensembles provided decision rules which could be translated to provide additional prognostic values to existing clinical prediction models[10,15,17,19], with the additional benefit of incorporating relevant genetic information into clinical decision making for PwMS. Finally, modeling the continuous-time evolution of EDSS, increased the variance in disability progression that is genetically determined.

Declarations

Acknowledgements

Special thanks to the participants who made this study possible; and the AUSLONG/AUSIMMUNE investigators group for designing the study and for granting approval (H0010499) and access to the AUSLONG datasets.

The AUSLONG/AUSIMMUNE investigators group members are**:** RL (National Centre for Epidemiology and Population Health, Canberra), Keith Dear (Duke Kunshan University, Kunshan, China), A-LP and Terry Dwyer (Murdoch Childrens Research Institute, Melbourne, Australia), IvdM, LB, SSY, BVT, and Ingrid van der Mei (Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia), SB (School of Medicine, Griffith University, Gold Coast Campus, Australia), Trevor Kilpatrick (Centre for Neurosciences, Department of Anatomy and Neuroscience, University of Melbourne, Melbourne, Australia). David Williams and Jeanette Lechner-Scott (University of Newcastle, Newcastle, Australia), Cameron Shaw and Caron Chapman (Barwon Health, Geelong, Australia), Alan Coulthard (University of Queensland, Brisbane, Australia), Michael P Pender (The University of Queensland, Brisbane, Australia) and Patricia Valery (QIMR Berghofer Medical Research Institute, Brisbane, Australia).

Author contributions

BVT***** supervised the study; VFN performed data analysis and wrote the manuscript; YZ, PEM, IVM, JC, XL, AZ, SAB, ALP, SSY, and JLS, contributed edits and completed revisions of the manuscript draft. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Additional information

Additional data (R-Codes for EDSS Imputation, tables, and figures) are available as a web supplementary file.

Funding

This work was supported by the National Health and Medical Research Council of Australia [APP1127819, 1947180, 544922], Kate-Scott Memorial Scholarship (to Valery Fuh-Ngwa); Multiple Sclerosis Research Australia; National Health and Medical Research Council investigator grant L1 [GNT1173155] (to Yuan Zhou); Henry Baldwin Trust and the Medical Research Future Fund [EPCP000008] (to Jac Charlesworth); and Macquarie Foundation Multiple Sclerosis Research Australia Senior Clinical Research Fellowship (to Bruce V. Taylor).

References

1. Mandel, M., Mercier, F., Eckert, B., Chin, P. & Betensky, R. A. Estimating Time to Disease Progression Comparing Transition Models and Survival Methods-An Analysis of Multiple Sclerosis Data. Biometrics **69**, 225–234, doi:10.1111/biom.12002 (2013).
2. Zarghami, A., Li, Y., Claflin, S. B., Van Der Mei, I. & Taylor, B. V. Role of environmental factors in multiple sclerosis. Expert Review of Neurotherapeutics, 1–20, doi:10.1080/14737175.2021.1978843 (2021).
3. Consortium, M. S. G. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science **365**, eaav7188, doi:10.1126/science.aav7188 (2019).
4. Consortium, T. G. P. An integrated map of genetic variation from 1,092 human genomes. Nature **491**, 56–65, doi:10.1038/nature11632 (2012).
5. Jackson, K. C. *et al.* Genetic model of MS severity predicts future accumulation of disability. Ann. Hum. Genet. **84**, 1–10, doi:10.1111/ahg.12342 (2020).
6. Pan, G. *et al.* Role of genetic susceptibility variants in predicting clinical course in multiple sclerosis: a cohort study. Journal of Neurology, Neurosurgery & Psychiatry **87**, 1204–1211, doi:10.1136/jnnp-2016-313722 (2016).
7. Jensen, C. J. *et al.* Multiple Sclerosis Susceptibility-Associated SNPs Do Not Influence Disease Severity Measures in a Cohort of Australian MS Patients. PloS one **5**, e10003, doi:10.1371/journal.pone.0010003 (2010).
8. Lin, R. *et al.* Association between multiple sclerosis risk-associated SNPs and relapse and disability – a prospective cohort study. Multiple Sclerosis Journal **20**, 313–321, doi:10.1177/1352458513496882 (2014).
9. Fuh-Ngwa, V. *et al.* Developing a clinical-environmental-genotypic prognostic index for relapsing-onset multiple sclerosis and clinically isolated syndrome. Brain Communications, doi:10.1093/braincomms/fcab288 (2021).
10. Tommasin, S. *et al.* Machine learning classifier to identify clinical and radiological features relevant to disability progression in multiple sclerosis. Journal of Neurology, doi:10.1007/s00415-021-10605-7 (2021).
11. Ramanujam, R. *et al.* Accurate classification of secondary progression in multiple sclerosis using a decision tree. Multiple Sclerosis Journal **27**, 1240–1249, doi:10.1177/1352458520975323 (2021).
12. Pellegrini, F. *et al.* Predicting disability progression in multiple sclerosis: Insights from advanced statistical modeling. Multiple Sclerosis Journal **26**, 1828–1836, doi:10.1177/1352458519887343 (2020).
13. Baranzini, S. E. *et al.* Prognostic biomarkers of IFNb therapy in multiple sclerosis patients. Multiple Sclerosis Journal **21**, 894–904, doi:10.1177/1352458514555786 (2015).
14. Hajjem, A., Bellavance, F. & Larocque, D. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation **84**, 1313–1328, doi:10.1080/00949655.2012.741599 (2014).
15. Yperman, J. *et al.* Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis. BMC Neurology **20**, doi:10.1186/s12883-020-01672-w (2020).
16. Tommasin, S. *et al.* Machine learning classifier to identify clinical and radiological features relevant to disability progression in multiple sclerosis. Journal of Neurology **268**, 4834–4845, doi:10.1007/s00415-021-10605-7 (2021).

17. Pinto, M. F. *et al.* Prediction of disease progression and outcomes in multiple sclerosis with machine learning. Scientific reports **10**, doi:10.1038/s41598-020-78212-6 (2020).
18. Law, M. T. *et al.* Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression. Multiple Sclerosis Journal - Experimental, Translational and Clinical **5**, 205521731988598, doi:10.1177/2055217319885983 (2019).
19. De Brouwer, E. *et al.* Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression. Computer Methods and Programs in Biomedicine **208**, 106180, doi:10.1016/j.cmpb.2021.106180 (2021).
20. Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D. & McCoy, R. G. Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c. Journal of biomedical informatics **89**, 56–67, doi:10.1016/j.jbi.2018.09.001 (2019).
21. Deng, H. Interpreting tree ensembles with inTrees. International Journal of Data Science and Analytics **7**, 277–287, doi:10.1007/s41060-018-0144-8 (2019).
22. Caie, P. D., Dimitriou, N. & Arandjelović, O. in Chap. 8-*Artificial Intelligence and Deep Learning in Pathology* (ed Stanley Cohen) 149–173 (Elsevier, 2021).
23. Lucas, R. *et al.* Observational analytic studies in multiple sclerosis: controlling bias through study design and conduct. The Australian Multicentre Study of Environment and Immune Function. Multiple Sclerosis Journal **13**, 827–839, doi:10.1177/1352458507077174 (2007).
24. World Medical Association Declaration of Helsinki. JAMA **310**, 2191, doi:10.1001/jama.2013.281053 (2013).
25. Thompson, A. J. *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet Neurology **17**, 162–173, doi:10.1016/s1474-4422(17)30470-2 (2018).
26. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Plos Genet **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).
27. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. Nature Protocols **9**, 2643–2662, doi:10.1038/nprot.2014.174 (2014).
28. Nicole, Rizopoulos, D. & Emmanuel. JointAI: Joint Analysis and Imputation of Incomplete Data in R. *arXiv pre-print server*, doi:None arxiv:1907.10867 (2020).
29. Voskuhl, R. R. *et al.* Sex differences in brain atrophy in multiple sclerosis. Biology of Sex Differences **11**, doi:10.1186/s13293-020-00326-3 (2020).
30. Ribbons, K. A. *et al.* Male Sex Is Independently Associated with Faster Disability Accumulation in Relapse-Onset MS but Not in Primary Progressive MS. PloS one **10**, e0122686, doi:10.1371/journal.pone.0122686 (2015).
31. Koch-Henriksen, N. & Sørensen, P. S. The changing demographic pattern of multiple sclerosis epidemiology. The Lancet Neurology **9**, 520–532, doi:https://doi.org/10.1016/S1474-4422(10)70064-8 (2010).
32. Ramagopalan, S. V. *et al.* Sex ratio of multiple sclerosis and clinical phenotype. European Journal of Neurology **17**, 634–637, doi:10.1111/j.1468-1331.2009.02850.x (2010).
33. Orton, S.-M. *et al.* Sex ratio of multiple sclerosis in Canada: a longitudinal study. The Lancet Neurology **5**, 932–936, doi:10.1016/S1474-4422(06)70581-6 (2006).
34. Mandel, M., Gauthier, S. A., Guttmann, C. R. G., Weiner, H. L. & Betensky, R. A. Estimating Time to Event From Longitudinal Categorical Data. J. Am. Stat. Assoc. **102**, 1254–1266, doi:10.1198/016214507000000059 (2007).
35. Mandel, M. & Betensky, R. A. Estimating time-to-event from longitudinal ordinal data using random-effects Markov models: application to multiple sclerosis progression. Biostatistics **9**, 750–764, doi:10.1093/biostatistics/kxn008 (2008).
36. Sweeting, M. J., Farewell, V. T. & De Angelis, D. Multi-state Markov models for disease progression in the presence of informative examination times: An application to hepatitis C. Statistics in Medicine **29**, 1161–1174, doi:10.1002/sim.3812 (2010).
37. de Wreede, L. C., Fiocco, M. & Putter, H. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. Computer Methods and Programs in Biomedicine **99**, 261–274, doi:https://doi.org/10.1016/j.cmpb.2010.01.001 (2010).
38. Goeman, J. J. L1 Penalized Estimation in the Cox Proportional Hazards Model. Biom. J. **52**, 70–84, doi:10.1002/bimj.200900028 (2010).
39. Ibrahim, M. A. Quantile regression in heteroscedastic varying coefficient models: testing and variable selection.

40. Sauerbrei, W. & Royston, P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society: Series A (Statistics in Society) **162**, 71–94, doi:10.1111/1467-985x.00122 (1999).
41. Royston, P. & Altman, D. G. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology **13**, 33, doi:10.1186/1471-2288-13-33 (2013).
42. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. **8**, doi:10.1038/s41467-017-01261-5 (2017).
43. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research **47**, D886-D894, doi:10.1093/nar/gky1016 (2019).
44. Zhou, Y. *et al.* Variation within MBP gene predicts disease course in multiple sclerosis. Brain Behav **7**, e00670, doi:10.1002/brb3.670 (2017).
45. Sadovnick, A. D. *et al.* Genetic modifiers of multiple sclerosis progression, severity and onset. Clinical immunology (Orlando, Fla.) **180**, 100–105, doi:10.1016/j.clim.2017.05.009 (2017).
46. Jokubaitis, V. G. & Butzkueven, H. A genetic basis for multiple sclerosis severity: Red herring or real? Molecular and cellular probes **30**, 357–365 (2016).
47. Shams, H. *et al.* Polygenic risk score association with multiple sclerosis susceptibility and phenotype in Europeans. Brain, doi:10.1093/brain/awac092 (2022).

Figures

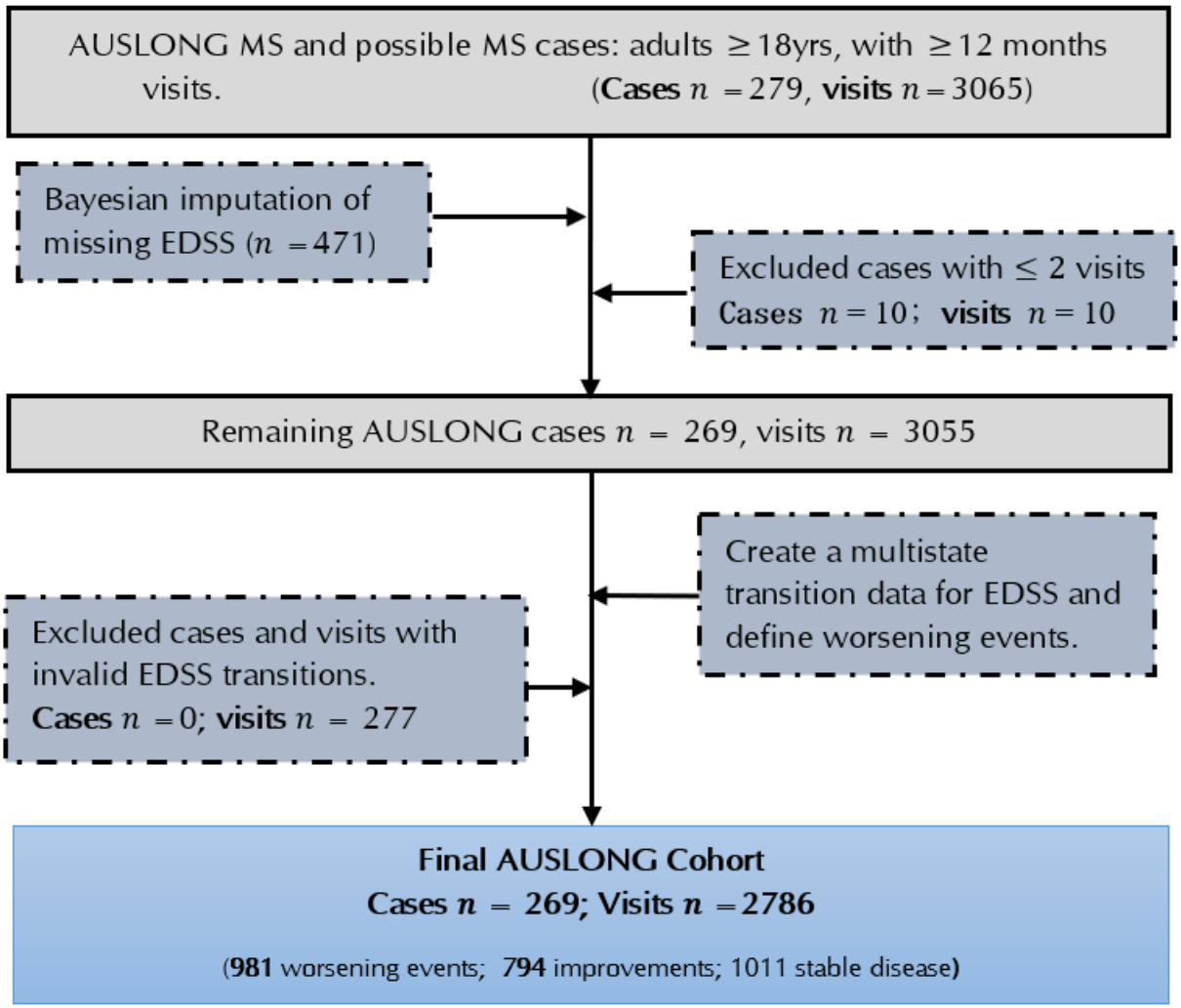

Figure 1

A flow chart of AUSLONG data extraction and case selection criteria.

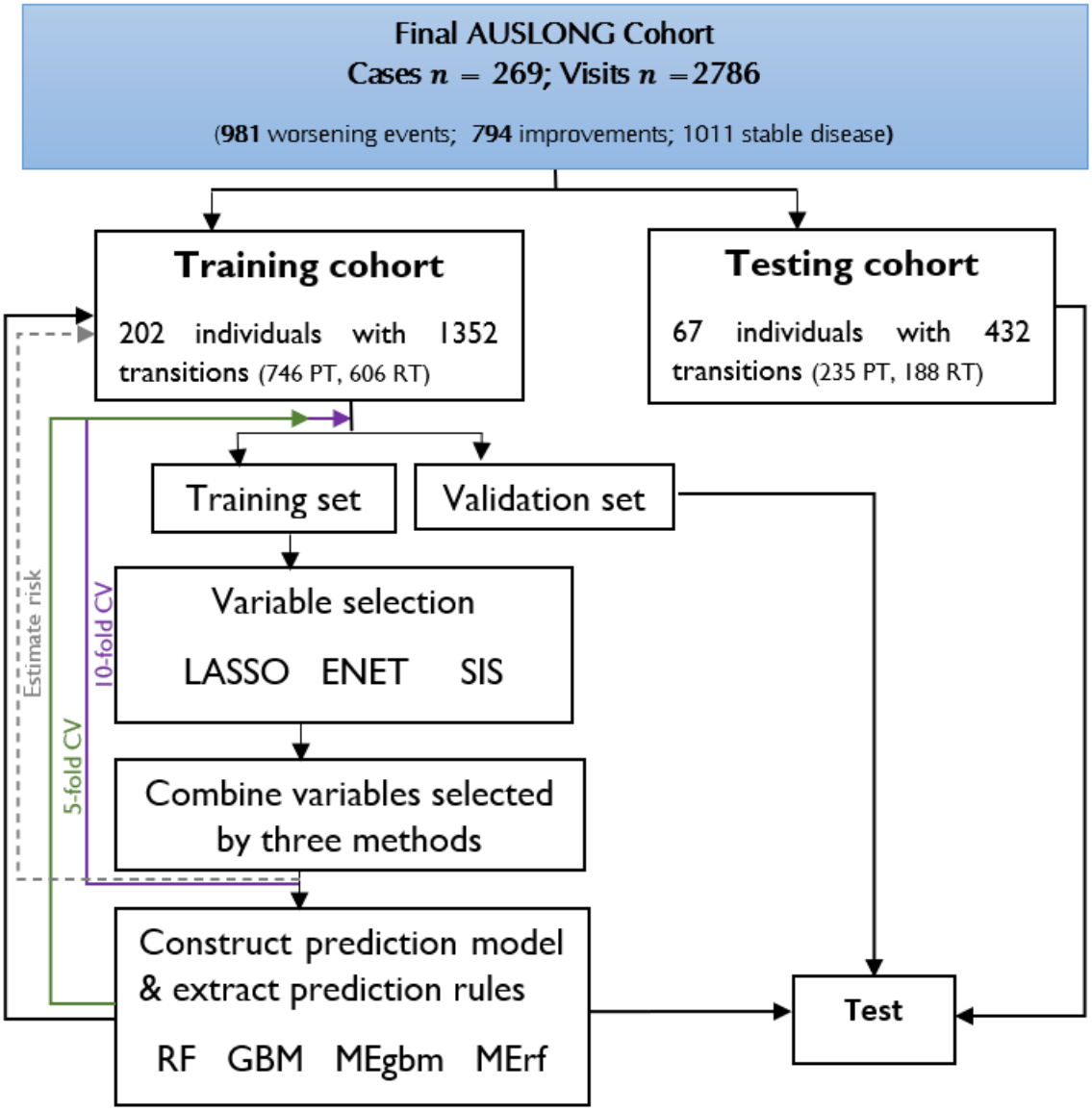

Figure 2

Outline of ensemble learning, and genetic risk prediction model construction. PT: progressive transitions ("worsening" events); RT: regressive transitions ("improved" events); ST: stable-state transitions ("sustained disability" or stable disease).

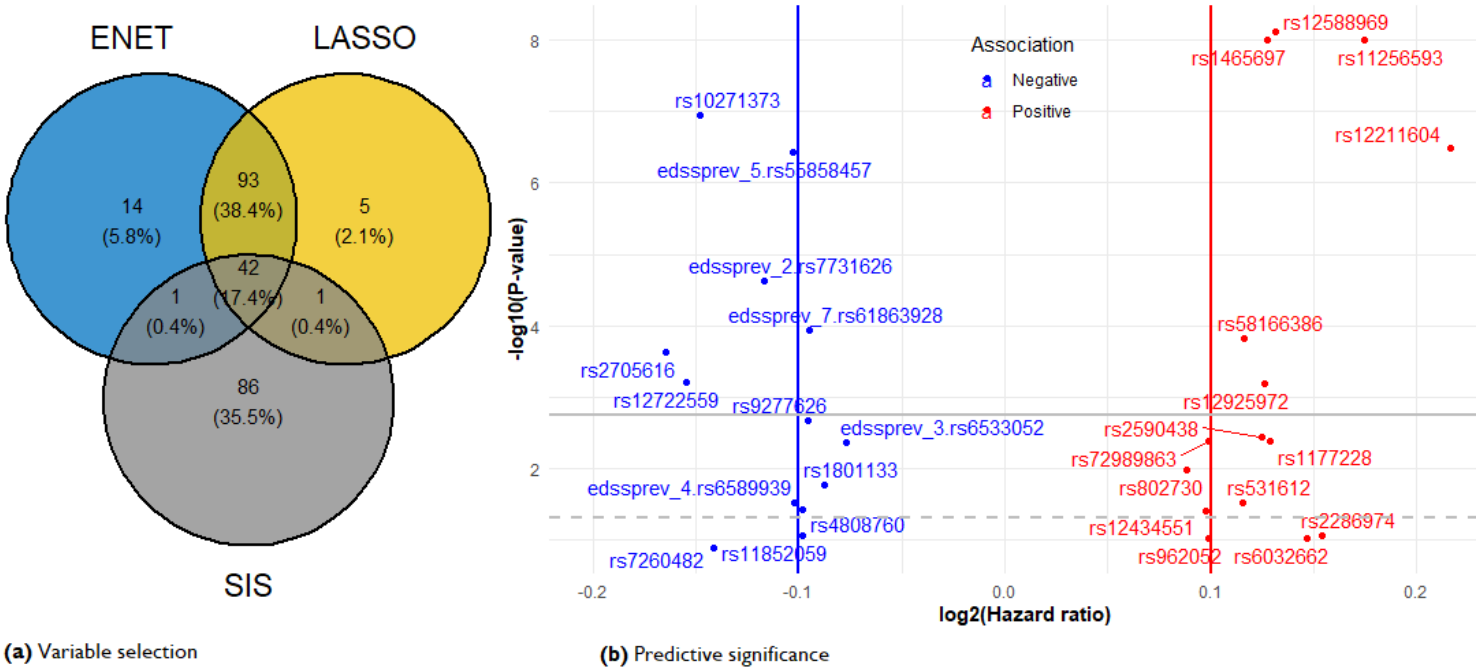

Figure 3

(a): Number of genetic variations shared among 10-fold CV by variable selection methods. (b): Volcano plot significance of SNPs in the multivariate prognostic model. Statistical significance at the multivariate cut-off level is −log10(α=0.05) = 1.30 (grey dash-lines), family-wise error rate at −log10(α =0.05/28) = 2.75 (grey solid-lines).

Figure 4

Predicting future disability progression. Correlation between the observed and predicted probability of worsening events stratified by MS phenotype in the training (a) and test cohort (b).

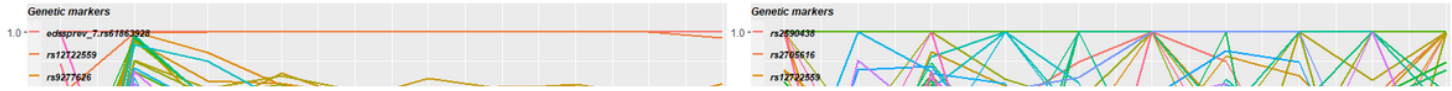

Figure 5

SNP ranking: Relative importance of genetic markers in predicting future worsening of disability over time.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialVFN.pdf