Diagnosing migraine from genome-wide genotype data: a machine learning analysis

Antonios Danelakis,[1,2] Tjaša Kumelj,[1,3] Bendik S. Winsvold,[1,4,5,6] Marte Helene Bjørk,[1,7,8] Parashkev Nachev,[9] Manjit Matharu,[1,10] Dominic Giles,[9] International Headache Genetic Consortium, Erling Tronvik,[1,3,11] Helge Langseth[1,2] and Anker Stubberud[1,3]

Abstract

Migraine has an assumed polygenic basis, but the genetic risk variants identified in genome-wide association studies only explain a proportion of the heritability. We aimed to develop machine learning models, capturing non-additive and interactive effects, to address the missing heritability.

This was a cross-sectional population-based study of participants in the second and third Trøndelag Health Study. Individuals underwent genome-wide genotyping and were phenotyped based on validated modified criteria of the International Classification of Headache Disorders. Four datasets of increasing number of genetic variants were created using different thresholds of linkage disequilibrium and univariate genome-wide associated p-values. A series of machine learning and deep learning methods were optimized and evaluated. The genotype tools PLINK and LDPred2 were used for polygenic risk scoring. Models were trained on a partition of the dataset and tested in a hold-out set. The area under the receiver operating characteristics curve was used as the primary scoring metric. Classification by machine learning was statistically compared to that of polygenic risk scoring. Finally, we explored the biological functions of the variants unique to the machine learning approach.

43,197 individuals (51% women), with a mean age of 54.6 years, were included in the modelling. A light gradient boosting machine performed best for the three smallest datasets (108, 7,771 and 7,840 variants), all with hold-out test set area under curve at 0.63. A multinomial naïve Bayes model performed best in the largest dataset (140,467 variants) with a hold-out test set area under curve of 0.62. The models were statistically significantly superior to polygenic risk scoring (area under curve 0.52 to 0.59) for all the datasets ($p<0.001$ to $p=0.02$). Machine learning identified

many of the same genes and pathways identified in genome-wide association studies, but also several unique pathways, mainly related to signal transduction and neurological function. Interestingly, pathways related to botulinum toxins, and pathways related to the calcitonin gene-related peptide receptor also emerged.

This study suggests that migraine may follow a non-additive and interactive genetic causal structure, potentially best captured by complex machine learning models. Such structure may be concealed where the data dimensionality (high number of genetic variants) is insufficiently supported by the scale of available data, leaving a misleading impression of purely additive effects. Future machine learning models using substantially larger sample sizes could harness both the additive and the interactive effects, enhancing precision and offering deeper understanding of genetic interactions underlying migraine.

Author affiliations:

1 NorHead Norwegian Centre for Headache Research, NTNU Norwegian University of Science and Technology, 7030, Trondheim, Norway

2 Department of Computer Science, NTNU Norwegian University of Science and Technology, 7034, Trondheim, Norway

3 Department of Neuromedicine and Movement Science, NTNU Norwegian University of Science and Technology, 7030, Trondheim, Norway

4 Department of Research and Innovation, Division of Clinical Neuroscience, Oslo University Hospital, 0450, Oslo, Norway

5 Department of Neurology, Oslo University Hospital, 0372, Oslo, Norway

6 HUNT Center for Molecular and Clinical Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, NTNU Norwegian University of Science and Technology, 7030, Trondheim, Norway

7 Department of Clinical Medicine, University of Bergen, 5021, Bergen, Norway

8 Department of Neurology, Haukeland University Hospital, 5053, Bergen, Norway

9 High Dimensional Neurology Group, UCL Institute of Neurology, University College London, London WC1N 3BG, UK

10 Headache and Facial Pain Group, UCL Institute of Neurology and National Hospital for Neurology and Neurosurgery, London WC1N 3BG, UK

11 Neuroclinic, St Olav University Hospital, 7030, Trondheim, Norway

Correspondence to: Anker Stubberud

NTNU, Faculty of Medicine and Health Sciences, N-7491 Trondheim, Norway

E-mail: anker.stubberud@ntnu.no

Running title: Diagnosing migraine from genotype data

Keywords: artificial intelligence; gradient boosting; headache; genetics; epistasis; HUNT

Introduction

Migraine is a common primary headache disorder with a substantial global disease burden.[1] The global prevalence is estimated to 14%,[2] and it is ranked second among causes of disability, and first among women under 50 years of age.[3] Migraine is characterized by recurring attacks of intense, often unilateral and pulsating headaches, accompanied by nausea, vomiting, and sensitivity to light and sound.[4] In up to a third of individuals the attacks are at times preceded by transient focal neurological aura symptoms, most commonly visual or sensory.

The etiology of migraine is complex and incompletely understood. Inheritance has long been recognized as important, as migraine tends to cluster in families.[5,6] Twin studies have confirmed consistently higher concordance rates of migraine in monozygotic twins versus dizygotic twins,[7] with an estimated heritability of around 50%.[8]

The largest genome-wide association study (GWAS) meta-analysis identified 123 migraine risk loci.[9] Other GWAS have identified similar and other risk variants.[10,11] Yet, the sum of the risk variants does not explain the full heritability of migraine.[12] Indeed, it was estimated that the 123

risk loci from the 2022 GWAS only explain 11.2% of the heritability. The missing heritability—defined as the gap between heritability estimates from twin studies and the heritability explained by the identified genetic variants—may be attributable to at least two factors.[13] First, there are likely many small-effect-variants that increase the risk of migraine but fail to reach the significance level required in GWAS. Second, there may be epistatic interactions, where the effect of one gene is modified by other genes complicating the genetic architecture. This results in an overall effect that is not merely the sum of each gene's contribution (additive effects) but is instead driven by the combined influence of interacting genes (non-additive effects).[14]

Polygenic risk scoring (PRS) can be used to estimate the additive risk of complex traits based on the sum of all risk alleles carried by an individual.[15] This summing across variants assumes an additive genetic architecture, with independence of risk variants,[15] and does not take into account any gene-gene or gene-environment interactions.[16] Such an approach is not suited to explain any interactive genetic factors contributing to the missing heritability.

Therefore, implementing a model that accounts for interactive effects–in addition to additive effects–could distinguish individuals with migraine from headache-free controls using genotype data with better precision. Importantly, it could also help explain the missing heritability and increase our understanding of the genetic architecture of migraine. We hypothesized that complex, high-dimensional machine learning models that can handle a large number of input variables while preserving covariate interactions may address the shortcomings of PRS.

The objectives of this study were to (1) estimate the accuracy of machine learning in distinguishing migraine from genome-wide genotype data; (2) compare the diagnostic accuracy of machine learning models with PRS across increasing dimensionalities (increasing number of genetic variants) of genetic input data, and (3) evaluate possible biological mechanisms of genes and interactions identified through machine learning modelling.

Materials and methods

Data sources and data materials

This was a cross-sectional population-based machine learning analysis of genome-wide genotype data for classifying individuals with migraine versus headache-free controls. The methods for

acquiring genotype data and phenotype assignment were similar to those reported in previous studies of the same health survey and biobank data material.[17,18]

The Trøndelag health study

The Trøndelag Health Study (HUNT) is a large, population-based cohort study from Trøndelag county in Norway that has been carried out in four waves (HUNT1 to HUNT4).[19] All inhabitants aged twenty years or older living in the county were invited to participate. Participation was based on informed, written consent, and the study was approved by the Regional Committee for Medical and Health Research (#2015/576/REK Midt and #2014/144/REK Midt). Data was collected through questionnaires and clinical examinations. DNA from whole blood was collected in HUNT2 (1995–1997) and HUNT3 (2006–2008). Questionnaire data for phenotype assignment was collected in HUNT2 and HUNT3.

Genotyping

Genotyping of HUNT2 and HUNT3 participants was performed at the Genomics-Core Facility at the NTNU Norwegian University of Science and Technology. Three different versions of the Illumina HumanCoreExome microarray (Illumina HumanCoreExome12 v.1.0, HumanCoreExome12 v.1.1 and HumanCoreExome24 with custom content) were used. The quality control and imputation has been described in detail elsewhere.[20] In brief, after rigorous quality control, genotypes were imputed using a customized reference panel consisting of the Haplotype Reference consortium release 1.1.. Finally, variants with imputation quality r^2 <0.3 were excluded.

Phenotype assignment

A diagnosis of migraine was assessed using a modified version of the International Classification of Headache Disorders[21,22] based on questionnaires in HUNT2 and HUNT3. Participants were asked whether they had suffered from headache during the last 12 months, and those who answered "yes" were classified as headache sufferers, while those who answered "no" constituted the control group of headache-free individuals. Those answering "yes" were subsequently asked questions about their headache to assess whether they fulfilled criteria for migraine or not. Those fulfilling the criteria for migraine were classified as migraine cases in this study. This method of phenotype

assignment is reported in detail elsewhere and has been validated through clinical interviews by a headache neurologist.[23,24]

Genome-wide association study data and dataset creation

From the largest migraine GWAS meta-analysis, which was based on 102,084 cases with migraine and 771,257 controls,[9] we acquired summary statistics for the lead variants of the 123 identified migraine risk loci, and the 8,117 migraine variants that reached the genome-wide significance threshold of $p<5*10^{-8}$. Because HUNT participants were part of the GWAS meta-analysis,[9] a reverse meta-analysis was conducted to derive a new beta coefficient and standard error for each variant after excluding individuals from the HUNT study. Using the recalculated beta and standard error, updated p-values for the migraine association were obtained by calculating the cumulative density function of a normal distribution, with a mean of 0.0 and a standard deviation of 1.0. This method of reverse meta-analysis allowed us to re-calculate the summary statistics without the influenced by HUNT individuals, in turn allowing machine learning and PRS classification of the "unseen" HUNT samples. The re-calculated summary statistics were used to create dataset 1 and 2 (see below).

In addition to the 2022 GWAS meta-analysis[9] summary statistics for significant variants, we used the complete, genome-wide summary statistics from this meta-analysis after excluding 23andMe, owing to data availability.[9] A reverse meta-analysis method, as described above, was again used to remove the influence of the HUNT individuals. These summary statistics were used to create dataset 3 and 4 (see below).

To compare the diagnostic accuracy of machine learning versus PRS and evaluate the effect of the dimensionality of the genotype input data, four different datasets with increasing number of genetic variants were created:

(1) the linkage disequilibrium independent variants with an r^2 threshold of 0.1, reaching the genome-wide significance threshold of $p<5*10^{-8,}$ identified among the 8,117 genome-wide significant variants from the 2022 GWAS meta-analysis[9] after having performed reverse meta-analysis to remove HUNT individuals;

(2) the variants available in our dataset, among the 8,117 variants from the 2022 GWAS meta-analysis,[9] reaching the genome-wide significance level of $p<5*10^{-8}$ after having performed reverse meta-analysis to remove HUNT individuals;

(3) the variants reaching a significance level of $p<1*10^{-5}$ captured from the summary statistics of the 2022 GWAS meta-analysis[9] excluding 23andMe and after having performed reverse meta-analysis to remove HUNT individuals;

(4) the LD-independent variants, at an r^2 threshold of 0.1, from available variants in the summary statistics of the 2022 GWAS meta-analysis[9] excluding 23andMe and after having performed reverse meta-analysis to remove HUNT individuals.

Machine diagnostic modelling

The genotyped variants (where available) or imputed variants (dosages, i.e. a decimal number between 0 and 2 describing the probability of the imputation corresponding to a given allele combination) were used as input variables (features) for the models. The genetic variant dosages were one-hot-encoded (redefined as dummy variables) in datasets 1-3. Dataset 4, was not one-hot-encoded because its dimensionality (number of variables) was already significant and one-hot-encoding would three-fold the feature size, resulting in a problematically large feature-to-sample size ratio.[25] The phenotype assignment (migraine or headache-free) was used as the outcome (label). Figure 1 is a schematic of the study design and modelling strategy.

The data was split in a random stratified fashion into a training and a test set in a 9:1 ratio. The test set was the same for all machine learning and PRS models and was kept unseen until the final model evaluation.

A series of standard machine learning classification architectures were evaluated: logistic regression, least absolute shrinkage and selection operator, support vector machines, decision trees, k-nearest neighbors, naïve Bayes, random forest, gradient boosting methods, and ensemble methods. Owing to the substantial number of features for dataset 4, this data was trained in chunks with the following classifiers: perceptron, stochastic gradient descent, passive-aggressive classifier, and multinomial, Bernoulli, gaussian and complement naïve Bayes. Finally, deep learning architectures, including TabNet,[26] GenNet[27] and fDDN[28] (specifically to tackle the issue with input dimensionality surpassing sample size), were evaluated. GenNet is a deep learning

network that preprocesses data based on genetic annotation. FDNN is a deep learning network specifically developed to handle cases with a large feature-to-sample ratio.

Model hyperparameters were optimized using Optuna[29] with 500 trials. Models were trained on the training set and performance was continuously evaluated with 10-fold cross-validation. For the largest dataset a single train/validate split was used owing to extensive compute time. The area under the receiver operating characteristics curve (AUC) was used as a scoring metric for training and optimizing the models. The mean AUC and it's standard deviation (SD) were calculated across the ten training folds to summarize the model's training performance and the variability between folds. In addition to AUC, we calculated accuracy, precision, recall and F1-score. The precision is the proportion of those classified as migraine that indeed have migraine and is identical with the positive predictive value. The recall is the proportion of those that have migraine that were classified as migraine and is the same as sensitivity. The F1-score is a compound metric of precision and recall. The top performing model for each of the four datasets was finally applied on the test set to quantify out-of-sample performance, calculating AUC with 95% confidence intervals (CI). All machine learning analyses were conducted using Python 3.10 (Python Software Foundation) with open-source packages (Supplementary table 1).

Sample characteristics, demographics and phenotype assignment were statistically described as proportions for dichotomous variables and means with SD for continuous variables.

Sensitivity analysis of relatedness

To estimate the influence of relatedness of individuals, we conducted a sensitivity analysis on unrelated individuals (up to 3rd degree relatedness was clumped), using the PLINK command *plink2 --bfile plinkFileName --king-cutoff 0.006*.

Sensitivity analysis of feature dimensionality

Post-hoc, a series of intermediate datasets with number of variants between dataset 3 and dataset 4 were created to better elucidate the impact of feature dimensionality and model complexity. These datasets were created by changing the linkage disequilibrium threshold. Each of the intermediate datasets were used to train the best simple additive machine learning model, and the best complex machine learning model.

Polygenic risk scoring

Two methods were employed for calculating PRS: PLINK,[30] and LDpred2.[31] PLINK extracts the PRS based on the clumping and thresholding approach. This minimizes the risk of overrepresenting certain genomic regions due to high linkage disequilibrium, using a linkage disequilibrium clumping correlation (r^2) cut-off value of 0.1 to determine which variants are considered too correlated. PRS was calculated for each subject using five different p-value thresholds (10^{-7}, 10^{-8}, 10^{-9}, 10^{-10} and 10^{-11}) representing the significance of association with the migraine label. The p-value threshold with the best model fit was used. Population structure was accounted for by incorporating ten principal components capturing ancestry-related differences as covariates to make the findings more reliable across diverse groups. Finally, the PRS were normalized and using a 0.5 decision threshold is used to distinguish cases and controls. LDpred2 also performs linkage disequilibrium clumping and accounts for population stratification using 10 principal components, similar to PLINK. Thereafter a logistic regression model is trained to achieve the best fit of the PRS on the data. Finally, the trained regression model is used to classify the phenotype. The dataset train/test split used for hold-out test set evaluation of the PRS approaches was identical to that of the machine learning models.

Comparison of machine diagnostics with polygenic risk scoring

To compare the diagnostic performance of machine learning and PRS, the test set AUCs were compared statistically. The null hypothesis criterion was tested by performing the Wilcoxon nonparametric test of independent samples.[32] The statistical significance threshold was set at 0.05.

Model explainability

Using the top performing model, we constructed calibration plots to check how accurately the model's classification matched the actual migraine outcomes. We also constructed probability density curves for both the machine learning models and PRS to visualize the separability of cases and controls. For the top performing machine learning model for each of the four datasets, we calculated Shapley values. For each dataset, variants were ordered by Shapley values from highest to lowest, and compared to the GWAS meta-analysis.[9] We also constructed a SHAP (Shapley Additive exPlanations) summary plots to visualize the relative contribution of increasing genotype input dimensionality.[33] SHAP is a framework utilizing Shapley values to explain machine learning

model predictions. SHAP assigns each feature an importance value, which enables interpretation of how much the feature contributes towards the prediction.

Gene annotation and pathway enrichment

To perform gene annotation and pathway enrichment analyses we identified the most influential variants in the top performing machine learning models and the lead variants identified in the GWAS meta-analysis. For fair comparison, the 123 most important variants were selected. In cases where several variants were considered equally important, so that the total number exceeded 123, all those of equal importance were used. Feature importances were extracted from the models and prioritized by their importance, providing an ordered list of each variant's contribution to the model.

Annotations of variants to genes were based on the proximity method that maps a genetic variant with its nearest gene (or to each of the genes it directly overlaps), using SNPnexus[34] with EnsemblDB[35] as a mapping reference.

Next, annotations of genes to pathways were preformed using the Reactome Pathway Database,[36] as implemented in SNPnexus. Pathway enrichment p-values were adjusted for multiple testing using the Benjamini-Hochberg method to control the false discovery rate[36]. The crude significance threshold was set at 0.05, while the false discovery rate threshold was set to 0.1. The same annotation and pathway enrichment approach was used for both the variants identified through GWAS and the variants identified through machine learning.

Results

Sample characteristics

Demographics and phenotype assignment

43,197 individuals with available genotype and phenotype data were included in the analyses. Supplementary Fig. 1 is a flow-chart of the study population. 10,286 individuals (24%) were classified as having migraine and 32,911 (76%) were classified as headache-free controls. Among those with migraine, 7,225 (70%) were women, and among the headache-free controls 15,088 (46%) were women. The mean age of the overall population was 54.6 (SD=17.3). The mean age

of the migraine cases and the headache-free controls was 46.8 (SD=14.0), and 57.1 (SD=17.5), respectively. All participants were of European ancestry. The distributions of individuals with migraine and headache-free controls were similar in the training, validation and test splits. In the training set, 8,322 (24%) had migraine and 26,667 (76%) were headache-free; in the validation set for dataset 4, 925 (24%) had migraine and 2,963 (76%) were headache-free; and in the test set, 1,039 (24%) had migraine and 3,281 (76%) were headache-free. Previous clinical validations of the phenotype assignment found that in HUNT2, the sensitivity was 69% and specificity 89% for migraine.[23] In HUNT3, the sensitivity was 67% and specificity was 96%.[24]

Genotype data

In HUNT2 and HUNT3, 71,680 individuals were genotyped. After quality control and imputation, a total of 9,832,846 variants were available for the 43,197 individuals included in the analysis.

In the reverse meta-analysis procedure, the influence of 7,801 cases and 32,423 controls from the HUNT study was removed in the creation of datasets 1-4, and the influence of 53,109 cases and 230,876 controls from 23andMe was removed in the creation of datasets 3 and 4. Thus 94,283 cases and 738,834 controls were used to calculate summary statistics for dataset 1 and 2 and 41,174 cases and 507,958 controls were used to calculate summary statistics for dataset 3 and 4. Of note, 1,395 migraine cases and 1,011 controls from HUNT could not be removed from the summary statistic calculation as they were part of a previous meta-analysis[37] already included in the 2022 GWAS meta-analysis.[9]

After quality control, imputation, the reverse meta-analysis procedure, calculation of summary statistics, linkage disequilibrium pruning and pruning based on p-values for variant association the number of variants in the four datasets were:

(1) dataset 1: 108variant;

(2) dataset 2: 7,771 variants;

(3) dataset 3: 7,840 variants;

(4) dataset 4: 140,467 variants.

Machine diagnostic performance

For the first three datasets (108, 7,771 and 7,840 variants) the top performing model was the light gradient boosting machine classifier, with cross-validated AUCs between 0.64 and 0.65, and cross-validated accuracies between 0.60 and 0.62 (Table 1). The hold-out test set AUC was 0.63 (95% CI: 0.61-0.65), 0.63 (95% CI: 0.61-0.65) and 0.63 (95% CI: 0.62-0.66) for the datasets with 108, 7,771 and 7,804 variants, respectively. The corresponding hold-out test accuracies were 0.60, 0.60 and 0.61. The hold-out test set precision ranged from 0.59 to 0.60, recall ranged from 0.59 to 0.60 and the F1-score ranged from 0.55 to 0.57 (Table 1)

In the largest dataset, containing 140,467 variants, the top performing model was the multinomial naïve Bayes classifier which achieved a validation set AUC of 0.62 and an accuracy of 0.57. The hold-out test set AUC was 0.62 (95% CI: 0.60-0.64) and the accuracy was 0.58. Test set precision, recall and F1-score was 0.64, 0.56 and 0.43, respectively. Table 1 and Figure 2 provide additional training and test performance metrics for the models. Supplementary table 2 provides all the out-of-sample and training experimental results for the best models of each learning approach for every dataset.

Relatedness sensitivity analysis

3,567 migraine cases and 10,417 controls were unrelated and included in the relatedness sensitivity analysis. The mean cross-validated training AUC for the relatedness sensitivity analysis ranged from 0.62 to 0.63 for all four datasets. The corresponding test set AUCs ranged from 0.61 to 0.63. Supplementary table 3 outlines all performance metrics for the sensitivity analysis.

Feature dimensionality sensitivity analysis

Five intermediate datasets with 19,473, 57,965, 71,188, 93,237 and 114,179 variants were created. With increasing feature dimensionality (i.e., higher number of variants) the performance of the complex models increased until suddenly reaching a performance drop, whereas the performance of simpler additive models such as the multinomial naive Bayes increased steadily until reaching a plateau (Figure 3 and Supplementary table 2). The light gradient boosting machine peaked at 92,237 variants with a training and test set AUC of 0.66.

Polygenic risk scores

The test set PRS AUCs using PLINK were 0.52 (95% CI 0.50-0.54) for the dataset with 108 variants, 0.52 (95% CI 0.51-0.55) for the dataset with 7,771 variants, 0.53 (95% CI 0.51-0.55) for the dataset with 7,840 variants, and 0.53 (95% CI 0.52-0.56) for the dataset with 140,467 variants. The corresponding test set PRS AUCs using LDpred2 were 0.55 (95% CI 0.53-0.57), 0.56 (95% CI 0.54-0.58), 0.59 (95% CI 0.57-0.61), and 0.59 (95% CI 0.57-0.61), respectively.

Comparison of machine learning and polygenic risk scores

The machine learning models outperformed PRS in all four datasets (Table 2). The difference in AUCs was most pronounced for datasets 1 through 3 ($P < 0.001$), and slightly less pronounced for the largest dataset ($P = 0.02$). Figure 3 illustrates the impact of feature dimensionality on performance for both machine learning and PRS.

Model explainability

Figure 4 visualizes the SHAP values for the top performing machine learning models for each dataset. These figures demonstrate that datasets 1 through 3 benefits from a model that may capture non-additive effects, whereas this advantage is lost in the largest dataset in favor of an additive probabilistic architecture. After ordering the variants by Shapley values, the top 123 variants were compared to those in the 2022 GWAS meta-analysis.[9] As expected, all 108 variants in dataset 1 were identified among the 123 from the GWAS meta-analysis. For the larger datasets, the number of common variants among the 123 most important were 13, eight and none for dataset 2, 3 and 4, respectively (Supplementary table 4). Supplementary Fig. 2 shows the probability distribution plot and calibration plots for the top performing machine learning model and PRS. Cases and controls showed largely overlapping prediction probability density plots, however more so for PRS as compared to the machine learning models.

Gene annotation and pathway enrichment

All 108 variants for dataset 1, the top 184 variants for dataset 2, the top 123 variants for dataset 3 and the top 1018 variants for dataset 4 were used for gene annotations and pathway enrichment. In both dataset 2 and dataset 4, several variants were considered equally important, thus these numbers exceeded 123 (184 and 1018, respectively). Table 3 and Figure 5 details the annotated

genes and enriched pathways found to be common with those identified in the GWAS meta-analysis and those unique to the machine learning models.

Gene comparison analysis resulted in the identification of 55 genes that were unique to the best interactive machine learning model (light gradient boosting machine in dataset 2), and 1002 genes that were unique to the best probabilistic/additive machine learning model (multinomial naive Bayes in dataset 4). Further, pathway enrichment analysis resulted in the identification of 91 and 590 additional pathways for the respective machine learning models. After pruning the identified pathways based on crude p-value threshold, 14 and 74 pathways were evaluated in detail (supplementary table 5 and 6). Among these 21 were considered significant after correcting for a false discovery rate of 0.1. The enriched pathways were primarily related to signal transduction and neurological function.

Discussion

In this study we found that machine learning outperforms PRS in distinguishing individuals with migraine from headache-free individuals based on genotype data. The best machine learning model achieved a hold-out test set AUC of 0.63 and the best PRS model achieved a hold-out test set AUC of 0.59. This is the first study to utilize machine learning to classify individuals with migraine and headache-free controls using genotype data.[38] Other studies aiming to classify headaches have mainly focused on clinical data, MRI data or other non-genetic paraclinical data.[38-40]

Though an AUC of 0.63 is modest in absolute terms, it is the increment in performance of flexible models over PRS that is critical here. Gene-environment interactions and the imprecision of single point disease prevalence set a comparatively low ceiling on maximal achievable performance from genotypic data alone.[41] But the substantial difference between twin study estimates of heritability and PRS performance[9,12] suggests genetic susceptibility may be mediated by wider and more complex genetic interactions than conventional PRS models are able to capture, as evidenced by the superior performance of the flexible models used in our study. Note that the comparison between machine learning and PRS was stacked in favor of PRS here, since the PRS was based on a meta-analysis of 94,283 migraine cases and 738,834 controls, while the machine learning models were based on the much smaller HUNT study population. Hence, if compared between datasets of

equal size, the magnitude of difference would be expected to be even larger in favour of machine learning.

Indeed, the findings from the three smallest datasets (108, 7,771 and 7,840 variants) support a non-additive and interactive genetic architecture for migraine, and can also explain why the machine learning approach outperforms PRS. Recall that the top performing models in these datasets was a light gradient boosting classifier, a model that can capture both non-linear relationships and interactions. Therefore, the observed superiority of these models over the purely additive PRS supports that the missing heritability may in part be attributed to non-additive effects such as gene-gene interactions. The notion that machine learning models can pick up non-linear and interactive effects of genotype is supported by empirical data from several other complex traits.[42] In an analysis of 34,702 individuals from eight U.S. cohorts, an extreme gradient boosting model was demonstrated to increase the variance explained, compared to PRS, between 22 and 100 percent for complex traits such as height, blood pressure and cholesterol levels.[42] That study supports our finding that complex machine learning models can capture non-linear and interactive effects also in migraine. It is further supported by several studies that have found that specific gene-gene interactions synergistically increase the susceptibility for migraine.[43-45]

We observe that the complex machine learning models show a small, but gradual, increase in performance with increasing genetic dimensionality before performance dramatically deteriorates beyond 93,237 variants (Figure 3). On the other hand, the simpler probabilistic naïve Bayes models show a steady increase in performance before reaching a plateau beyond 57,965 variants. These patterns can be explained as follows:

The machine learning models perform only slightly better in dataset 2 and 3 as compared to dataset 1, likely because information from the same relatively small set of loci is used across all three models. This is evident from the SHAP plot (figure 5) where the light gradient boosting seems to prioritize slightly less than a fifth of variants. Notably, dataset 3 used a higher p-value threshold for association, but was drawn from a smaller sample, likely resulting in identifying variants from the same set of loci as dataset 1 and 2.

When information from additional parts of the genome are incorporated in the models in the intermediate post-hoc analyses (recall that these datasets used increasingly higher r^2 cut-offs), it led to an increase in performance, before the sudden stall beyond 93,237 variants. This drop in

performance can be explained by overfitting when feature dimensionality and complexity surpass what can be supported by the available sample size. A rule of thumb states that there should be at a minimum 5-10 samples for each feature (or dimension).[46] However, despite performance initially increasing as the number of dimensions increases, beyond a certain dimensionality, the performance deteriorates.[46]

On the contrary, the relatively "simpler" multinomial naïve Bayes model, assuming an additive probabilistic architecture, similar to PRS, plateaus beyond 57,965 variants where additional small-effect-variants provide negligible additional performance. This plateauing is, as expected, also observed for PRS (figure 3). In summary, we argue that complex models capture non-additive effects as long as the feature to sample ratio is appropriate, beyond which the simpler models are favoured. This paradoxical phenomenon supports the second explanation for the "missing heritability", namely that there are many small-to-medium size variants that fail to reach the genome-wide significance threshold but have an impact in PRS and additive models such as naïve Bayes.

In this study, we identified several genes and pathways that seem to be unique for the machine learning approach. While the best model for the smallest dataset primarily identified genes and pathways already established in the GWAS meta-analysis, the complex models of dataset 2 and 3 resulted in several unique genes and pathways. The majority of the most important variants as identified by the Shapley analysis were also unique for datasets 2 and 3. The genes annotated to these variants were primarily enriched in pathways related to signal transduction and neurological function, which is biologically plausible for migraine. Dataset 4 with 140,167 variants and a probabilistic naïve Bayes model resulted in almost exclusively unique genes. This is likely due to the larger number of variants included in the annotation and pathway analysis, naturally leading to inclusion of a wider part of the genome and thus significantly more genes and pathways. Therefore, any biological interpretations from this dataset must be done with caution.

Interestingly, the overall best model (light gradient boosting machine in dataset 2) highlighted pathways related to calcitonin gene-like receptors, and the toxicity of botulinum toxin A, D, E and F. The calcitonin gene-related ligand and its receptor play an important role in migraine pathophysiology where they mediate trigeminovascular pain transmission and vasodilatory neurogenic inflammation.[47] They are also targets of several monoclonal antibodies that have

demonstrated effect in preventing migraine.[48] One of the risk loci identified in the 2022 GWAS meta-analysis contains the gene encoding calcitonin gene-related peptide itself, but not its receptor.[9] The identification of the receptor in this study suggests that both the ligand and receptor are relevant for the susceptibility of migraine. OnabotulinumtoxinA is a therapeutic agent used as preventive treatment for migraine.[49] Its mechanism of action is thought to be the inhibition of pro-inflammatory and excitatory neurotransmitters and neuropeptides from primary afferent nociceptive pain fibers in the head and neck that participate in the development of peripheral and central sensitization.[50] The pathway identified here involve the SV2A gene, encoding synaptic vesicle glycoprotein 2A, which has been shown to be the receptor for botulinum toxin A.[51] It is therefore conceivable that an upregulation of the receptor increase the susceptibility to both migraine and a treatment effect of OnabotulinumtoxinA.

The approach of complex genotype modelling has several potential downstream clinical implications. First, future models with improved performance could serve as an objective measure of migraine. Second, the modeling approach is transferrable and could prove a valuable risk scoring tool for other phenotypically diverse, idiopathic neurological traits of non-additive genetic architecture. Finally, further unraveling of the model architecture could help elucidate the underlying etiology and pathophysiology of migraine, paving the way for clinical and therapeutic markers.

We believe that complex models that can capture both interactive and additive effects will further improve classification by genotype, given a sufficiently large sample. The prerequisites for such models to be successful rely on sufficiently large sample sizes to allow complex modelling without overfitting; and the use of the right computational algorithms, such as non-linear machine learning models and deep neural networks. Future efforts to classify migraine by genotypic data adhering to these prerequisites are likely to outperform the classification performance of this study. Moreover, future studies should aim to incorporate demographic, phenotypic and other medical data that could further take advantage of important gene-environment and epigenetic factors that most likely partake in the migraine etiology.[12] Finally, it is important that future research efforts also aim to validate the models in out-of-sample cohorts to assess their generalizability.

Strengths and limitations

This paper has several strengths. First, the models are rigorously validated in a held-out unseen test set. The test set performances are faithful to the trained model suggesting that it is generalizable. Secondly, the developed models are compared to a validated standard, namely PRS, which establishes its robustness. Both strengths overcome challenges that are repeatedly cited as barriers to why machine learning fails to prove clinically useful in medicine.[38,52] Thirdly, the models are free from any apparent data leaks, contrary to what typically happens when the input for the classification models is the symptomatology of the migraine which is what determines the headache status, thus leading to overly optimistic classification results. Weaknesses of the study includes the moderate sensitivity of the phenotype assignment, although with near-perfect specificity—a potential classification bias. However, because migraine is the minority class, with lower sensitivity the class imbalance increases, thereby creating a more challenging classification task which ultimately leads to underestimation of the model precision. Another limitation is that we were not able to remove the influence of a few HUNT individuals in the calculation of summary statistics, which could have biased the models in favour of the dataset at hand. Nevertheless, this limitation is expected to increase the performance also of the PRS, hence this weakness does not invalidate the finding that machine learning outperforms PRS.

When comparing the machine learning and PRS, there are several strengths and weaknesses of both approaches that should be acknowledged. The most important strength of the machine learning models for the task at hand is the ability to capture non-additive and interactive effects. However, it comes at the cost of often high computational time and limited interpretability. PRS on the other hand, is a validated and commonly accepted method of assessing the risk of complex traits and is much less computationally expensive.[30] Still, it is limited to assessing additive genetic architectures, which likely is insufficient for migraine.[12]

Conclusion

In this study we demonstrate that machine learning outperforms PRS in distinguishing migraine from headache-free controls when using genome-wide genotype data and succeed in identifying new genes and pathways potentially implicated in the disease. Complex machine learning models significantly outperform PRS when the number of genetic variants are relatively low, supporting

a non-additive and interactive genetic architecture. However, this benefit diminishes with increasing input dimensionality in favour of additive effects. Our findings support both an additive and an interactive and non-additive genetic basis for migraine, validating the hypothesized explanations for the missing heritability. Future research investigating larger cohorts with complex models that capture both additive and interactive relationships could likely improve classification performance based on genotype.

Data availability

The minimum dataset required to replicate this work contains personal sensitive information and is not publicly available nor available upon request. The analytical code may be provided upon reasonable request.

Acknowledgements

The Trøndelag Health Study (HUNT) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology NTNU), Trøndelag County Council, Central Norway Regional Health Authority, and the Norwegian Institute of Public Health. The genotyping was financed by the National Institute of health (NIH), University of Michigan, The Norwegian Research council, and Central Norway Regional Health Authority and the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU). The genotype quality control and imputation has been conducted by the K.G. Jebsen center for genetic epidemiology, Department of public health and nursing, Faculty of medicine and health sciences, Norwegian University of Science and Technology (NTNU). References for HUNT: https://pubmed.ncbi.nlm.nih.gov/36777998/, https://pubmed.ncbi.nlm.nih.gov/22879362/, https://pubmed.ncbi.nlm.nih.gov/35578897/

Funding

The Research Council of Norway. The funder had no role in design, conduction or interpretation of the study.

Competing interests

Anker Stubberud has received lecture honoraria from TEVA. AS is share-holder and patent holder of Nordic Brain Tech AS and the Cerebri app.

Supplementary material

Supplementary material is available at *Brain* online.

Appendix 1

International Headache Genetics Consortium

Full details are provided in the Supplementary material.

Verneri Anttila, Ville Artto, Andrea C. Belin, Anna Bjornsdottir, Gyda Bjornsdottir, Dorret I. Boomsma, Sigrid Børte, Mona A. Chalmer, Daniel I. Chasman, Bru Cormand, Ester Cuenca-Leon, George Davey-Smith, Irene de Boer, Martin Dichgans, Tonu Esko, Tobias Freilinger, Padhraig Gormley, Lyn R. Griffiths, Eija Hämäläinen, Thomas F. Hansen, Aster V. E. Harder, Heidi Hautakangas, Marjo Hiekkala, Maria G. Hrafnsdottir, M. Arfan Ikram, Marjo-Riitta Järvelin, Risto Kajanne, Mikko Kallela, Jaakko Kaprio, Mari Kaunisto, Lisette J. A. Kogelman, Espen S. Kristoffersen, Christian Kubisch, Mitja Kurki, Tobias Kurth, Lenore Launer, Terho Lehtimäki, Davor Lessel, Lannie Ligthart, Sigurdur H. Magnusson, Rainer Malik, Bertram Müller-Myhsok, Carrie Northover, Dale R. Nyholt, Jes Olesen, Aarno Palotie, Priit Palta, Linda M Pedersen, Nancy Pedersen, Matti Pirinen, Danielle Posthuma, Patricia Pozo-Rosich, Alice Pressman, Olli Raitakari, Caroline Ran, Gudrun R. Sigurdardottir, Hreinn Stefansson, Kari Stefansson, Olafur A. Sveinsson, Gisela M. Terwindt, Thorgeir E. Thorgeirsson, Arn M. J. M. van den Maagdenberg, Cornelia van Duijn, Maija Wessman, Bendik S. Winsvold, John-Anker Zwart.

References

1. Stovner LJ, Nichols E, Steiner TJ, *et al.* Global, regional, and national burden of migraine and tension-type headache, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology.* 2018;17(11):954-976.
2. Stovner LJ, Hagen K, Linde M, Steiner TJ. The global prevalence of headache: an update, with analysis of the influences of methodological factors on prevalence estimates. *The journal of headache and pain.* 2022;23(1):34.
3. Steiner T, Stovner L, Jensen R, Uluduz D, Katsarava Z. Migraine remains second among the world's causes of disability, and first among young women: findings from GBD2019. *The journal of headache and pain.* 2020;21:137-141.
4. Headache Classification Committee of the International Headache Society (IHS) The International Classification of Headache Disorders, 3rd edition. *Cephalalgia.* 2018;38(1):1-211.
5. Merikangas KR, Risch NJ, Merikangas JR, Weissman MM, Kidd KK. Migraine and depression: association and familial transmission. *J Psychiatr Res.* 1988;22(2):119-129.
6. Russell MB, Olesen J. Increased familial risk and evidence of genetic factor in migraine. *British medical journal.* 1995;311(7004):541-545.
7. Honkasalo ML, Kaprio J, Winter T, Heikkilä K, Sillanpää M, Koskenvuo M. Migraine and concomitant symptoms among 8167 adult twin pairs. *Headache.* 1995;35(2):70-78.
8. Nielsen CS, Knudsen GP, Steingrímsdóttir Ó A. Twin studies of pain. *Clin Genet.* 2012;82(4):331-340.
9. Hautakangas H, Winsvold BS, Ruotsalainen SE, *et al.* Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles. *Nature genetics.* 2022;54(2):152-160.
10. Bjornsdottir G, Chalmer MA, Stefansdottir L, *et al.* Rare variants with large effects provide functional insights into the pathology of migraine subtypes, with and without aura. *Nature genetics.* 2023;55(11):1843-1853.

11. Choquet H, Yin J, Jacobson AS, *et al.* New and sex-specific migraine susceptibility loci identified from a multiethnic genome-wide meta-analysis. *Communications Biology.* 2021;4(1):864.

12. Grangeon L, Lange KS, Waliszewska-Prosół M, *et al.* Genetics of migraine: where are we now? *The journal of headache and pain.* 2023;24(1):12.

13. Manolio TA, Collins FS, Cox NJ, *et al.* Finding the missing heritability of complex diseases. *Nature.* 2009;461(7265):747-753.

14. Eising E, de Vries B, Ferrari MD, Terwindt GM, van den Maagdenberg AM. Pearls and pitfalls in genetic studies of migraine. *Cephalalgia.* 2013;33(8):614-625.

15. Lewis C, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine.* 2020;12(1):44.

16. Aschard H. A perspective on interaction effects in genetic association studies. *Genetic epidemiology.* 2016;40(8):678-688.

17. Brumpton BM, Graham S, Surakka I, *et al.* The HUNT study: A population-based cohort for genetic research. *Cell Genom.* 2022;2(10):100193.

18. Børte S, Zwart J-A, Skogholt AH, *et al.* Mitochondrial genome-wide association study of migraine – the HUNT Study. *Cephalalgia.* 2020;40(6):625-634.

19. Krokstad S, Langhammer A, Hveem K, *et al.* Cohort profile: the HUNT study, Norway. *International journal of epidemiology.* 2013;42(4):968-977.

20. Winsvold BS, Kitsos I, Thomas LF, *et al.* Genome-Wide Association Study of 2,093 Cases With Idiopathic Polyneuropathy and 445,256 Controls Identifies First Susceptibility Loci. *Front Neurol.* 2021;12:789093.

21. Hedache Classification Subcommitee of the International Headache Society. The International Classification of Headache Disorders: 2nd edition. *Cephalalgia.* 2004;24:9-160.

22. Headache Classification Committee of the International Headache Society. Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. *Cephalalgia.* 1988;8:1-96.

23. Hagen K, Zwart JA, Aamodt AH, *et al.* The validity of questionnaire-based diagnoses: the third Nord-Trøndelag Health Study 2006-2008. *The journal of headache and pain.* 2010;11(1):67-73.

24. Hagen K, Zwart JA, Vatten L, Stovner LJ, Bovim G. Head-HUNT: validity and reliability of a headache questionnaire in a large population-based study in Norway. *Cephalalgia.* 2000;20(4):244-251.

25. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics.* 2005;21(8):1509-1515.

26. Arik SÖ, Pfister T. Tabnet: Attentive interpretable tabular learning. In: *Proceedings of the AAAI conference on artificial intelligence.* 2021:6679-6687.

27. van Hilten A, Kushner SA, Kayser M, *et al.* GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Communications Biology.* 2021;4(1):1094.

28. Kong Y, Yu T. A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification. *Scientific Reports.* 2018;8(1):16477.

29. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.* 2019:2623-2631.

30. Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics.* 2007;81(3):559-575.

31. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics*. 2020;36(22-23):5424-5431.

32. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36.

33. Merrick L, Taly A. The explanation game: Explaining machine learning models using shapley values. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction.* Springer. 2020:17-38.

34. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic acids research.* 2020;48(1):185-192.

35. Hubbard TJ, Aken BL, Beal K, *et al.* Ensembl 2007. *Nucleic acids research.* 2007;35:610-617.

36. Jassal B, Matthews L, Viteri G, *et al.* The reactome pathway knowledgebase. *Nucleic acids research.* 2020;48(1):498-503.

37. Gormley P, Anttila V, Winsvold BS, *et al.* Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature genetics.* 2016;48(8):856-866.

38. Stubberud A, Langseth H, Nachev P, Matharu MS, Tronvik E. Artificial intelligence and headache. *Cephalalgia.* 2024;44(8):1-13.

39. Ihara K, Dumkrieger G, Zhang P, Takizawa T, Schwedt TJ, Chiang C-C. Application of Artificial Intelligence in the Headache Field. *Current pain and headache reports.* 2024:1-9.

40. Torrente A, Maccora S, Prinzi F, *et al.* The clinical relevance of artificial intelligence in migraine. *Brain Sciences.* 2024;14(1):85-99.

41. Yeh P-K, An Y-C, Hung K-S, Yang F-C. Influences of Genetic and Environmental Factors on Chronic Migraine: A Narrative Review. *Current Pain and Headache Reports.* 2024;28(4):169-180.

42. Elgart M, Lyons G, Romero-Brufau S, *et al.* Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications biology.* 2022;5(1):856-867.

43. Alves-Ferreira M, Quintas M, Sequeiros J, *et al.* A genetic interaction of NRXN2 with GABRE, SYT1 and CASK in migraine patients: a case-control study. *The journal of headache and pain.* 2021;22(1):57.

44. Quintas M, Neto JL, Pereira-Monteiro J, *et al.* Interaction between γ-aminobutyric acid A receptor genes: new evidence in migraine susceptibility. *PloS one.* 2013;8(9):e74087.

45. Gonçalves FM, Luizon MR, Speciali JG, Martins-Oliveira A, Dach F, Tanus-Santos JE. Interaction among nitric oxide (NO)-related genes in migraine susceptibility. *Molecular and cellular biochemistry.* 2012;370:183-189.

46. Zollanvari A, James AP, Sameni R. A theoretical analysis of the peaking phenomenon in classification. *Journal of Classification.* 2020;37:421-434.

47. Puledda F, Silva EM, Suwanlaong K, Goadsby PJ. Migraine: from pathophysiology to treatment. *Journal of neurology.* 2023;270(7):3654-3666.

48. Oliveira R, Gil-Gouveia R, Puledda F. CGRP-targeted medication in chronic migraine-systematic review. *The journal of headache and pain.* 2024;25(1):51.

49. Lanteri-Minet M, Ducros A, Francois C, Olewinska E, Nikodem M, Dupont-Benjamin L. Effectiveness of onabotulinumtoxinA (BOTOX®) for the preventive treatment of chronic migraine: A meta-analysis on 10 years of real-world data. *Cephalalgia.* 2022;42(14):1543-1564.

50. Burstein R, Blumenfeld AM, Silberstein SD, Manack Adams A, Brin MF. Mechanism of action of onabotulinumtoxinA in chronic migraine: a narrative review. *Headache: The Journal of Head and Face Pain.* 2020;60(7):1259-1272.

51. Dong M, Yeh F, Tepp WH, *et al.* SV2 is the protein receptor for botulinum neurotoxin A. *Science.* 2006;312(5773):592-596.

52. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nature medicine.* 2022;28(1):31-38.

Figure legends

Figure 1 Schematic overview of the study design. Among 43,197 individuals, 10,286 had migraine and 32,911 were headache-free controls. Four different datasets with an increasing number of genetic variants were used for distinguishing migraine vs. headache-free controls. These datasets were split in the same 9:1 ratio training and test sets. The training data was subsequently preprocessed, trained and optimized using 10-fold cross-validation. The best model for each dataset was evaluated on the test set.

Figure 2 Performance of the best machine learning models. Receiver operating characteristics curves for top performing machine learning models for each of the four datasets showing mean 10-fold cross-validated area under curve (blue line) ± 1 standard deviation (grey shaded area), and test set area under curve (orange line). (**A**) Dataset 1 with 108 variants. (**B**) Dataset 2 with 7,771 variants. (**C**) Dataset 4 with 7,840 variants. (**D**) Dataset 4 with 140,467 variants.

Figure 3 Impact of feature dimensionality. The hold-out test set area under curve (y-axis) is plotted against the number of variants included in the model (y-axis) for the best machine learning and polygenic risk scoring approaches. For each color, solid lines represent training performance and dotted lines represent test performance for a given modelling approach. Performance for the intermediate datasets (19,473 to 114,179 variants) were only calculated for the best non-linear complex machine learning approach (light gradient boosting) and the best simple additive model (multinomial naïve Bayes) as part of the post-hoc sensitivity analyses. Note that light gradient boosting increases in performance up to 93,237 variants before sharply dropping, indicating overfitting when the feature space exceeds a limit. Multinomial naïve Bayes, however, increases steadily before reaching a plateau beyond 57,965 variants. LightGBM: light gradient boosting machine. MNB: multnomial naïve Bayes.

Figure 4 SHAP summary plots. Plots illustrating the relative contribution of the included variants to the predictions for the best machine learning model for each dataset. The x-axes denote number of variants, the y-axes denote the absolute SHAP value on a logarithmic scale. (**A**) In dataset 1, all 108 variants contributed towards the prediction. (**B**) In dataset 2, 1,486 out of 7,771 variants contributed. (**C**) In dataset 3, 1,442 out of 7,840 variants contributed. In the two latter cases, a large majority of variants do not contribute to the prediction suggesting that the model omits the less important variants, however, still achieving higher accuracy than polygenic risk scoring suggesting that some non-additive effects between the contributing variants are captured. (**D**) In dataset 4, all 140,467 variants contribute but with small contribution each. This is due to the probabilistic additive architecture of the naive Bayes approach, more similar to polygenic risk scoring. SHAP: Shapley additive explanations.

Figure 5 Venn diagrams showing overlap of annotated genes and enriched pathways. (**A**) Overlap of annotated genes from variants identified in the genome-wide association study, the best complex model (light gradient boosting machine in dataset 2) and the best additive machine learning model (multinomial naïve Bayes in dataset 4). (**B**) Overlap of enriched pathways from genes and variants identified in the genome-wide association study, the best complex model (light gradient boosting machine in dataset 2) and the best additive machine learning model (multinomial naïve Bayes in dataset 4).

Table 1 Performance of best machine learning models

	AUC		Accuracy		Recall		Precision		F1-Score	
Dataset	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Dataset 1 (108 variants)[a]	0.64 ± 0.010	0.63	0.60 ± 0.009	0.60	0.59 ± 0.010	0.59	0.59 ± 0.010	0.59	0.55 ± 0.009	0.55
Dataset 2 (7771 variants)[b]	0.64 ± 0.010	0.63	0.62 ± 0.008	0.61	0.60 ± 0.009	0.59	0.60 ± 0.009	0.59	0.55 ± 0.009	0.55
Dataset 3 (7840 variants)[c]	0.65 ± 0.012	0.63	0.61 ± 0.009	0.62	0.60 ± 0.011	0.60	0.60 ± 0.011	0.60	0.58 ± 0.010	0.57
Dataset 4 (140 467 variants)[d]	0.62	0.62	0.59	0.58	0.57	0.56	0.65	0.64	0.45	0.43

For each scoring metric, the training set performance is presented as the mean of 10-fold cross validation (± standard deviation), except for dataset 4, in which only one train/validate split was evaluated. The test value is the performance of the trained model in the hold-out test set.
[a]Lead variants from risk loci identified in the 2022 genome-wide association study meta-analysis[10] (108 variants).
[b]All variants with p-value $<5 \times 10^{-8}$ identified from the 2022 genome-wide association study meta-analysis[10] summary statistics (7771 variants).
[c]All variants with a p-value $<1 \times 10^{-5}$ identified from the re-calculated summary statistics without 23andMe (7840 variants).
[d]All linkage disequilibrium independent variants among all genotyped variants (140 467 variants).

Table 2 Comparison of machine learning and polygenic risk scoring

Dataset	Best machine learning model AUC (95% CI)	PLINK AUC (95% CI)	LDpred2 AUC (95% CI)	Comparison
Dataset 1 (108 variants)[a]	0.63 (0.61–0.65)	0.52 (0.50–0.54	0.55 (0.53–0.57)	$P < 0.001$
Dataset 2 (7771 variants)[b]	0.63 (0.61–0.65)	0.52 (0.51–0.55)	0.56 (0.54–0.58)	$P < 0.001$
Dataset 3 (7840 variants)[c]	0.63 (0.62–0.66)	0.53 (0.51–0.55)	0.59 (0.57–0.61)	$P < 0.001$
Dataset 4 (140 467 variants)[d]	0.62 (0.60–0.64)	0.53 (0.52–0.56)	0.59 (0.57–0.61)	$P = 0.02$

In datasets 1, 2, and 3, a light gradient boosting machine performed best. In dataset 4, a multinomial naïve Bayes model performed best. The rightmost column is the result of a Wilcoxon nonparametric test comparing the best machine learning model and the best polygenic risk scoring approach.
[a]Lead variants from risk loci identified in the 2022 genome-wide association study meta-analysis[10] (108 variants).
[b]All variants with p-value $<5 \times 10^{-8}$ identified from the 2022 genome-wide association study meta-analysis[10] summary statistics (7771 variants).
[c]All variants with a p-value $<1 \times 10^{-5}$ identified from the re-calculated summary statistics without 23andMe (7840 variants).
[d]All linkage disequilibrium independent variants among all genotyped variants (140 467 variants).

Table 3 Comparison of identified genes and pathways

	Dataset 1 (108 variants)[a]	Dataset 2 (7771 variants)[b]	Dataset 3 (7840 variants)[c]	Dataset 4 (140 467 variants)[d]
Common genes, n/n (%)	73/105 (69.5)	60/115 (52.2)	38/95 (40.0)	8/1010 (0.8)
Genes unique to machine learning, n/n (%)	32/105 (30.5)	55/115 (47.8)	57/95 (60.0)	1002/1010 (99.2)
Common pathways, n/n (%)	228/254 (89.8)	139/230 (60.4)	127/226 (56.2)	200/790 (25.3)
Pathways unique to machine learning, n/n (%)	26/254 (10.2)	91/230 (39.6)	99/226 (43.8)	590/790 (74.7)

After gene annotation and pathway enrichment, the significant genes and pathways identified in the best machine learning models were compared to those identified in through a genome-wide association study approach.

[a]Lead variants from risk loci identified in the 2022 genome-wide association study meta-analysis[10] (108 variants).

[b]All variants with p-value $<5 \times 10^{-8}$ identified from the 2022 genome-wide association study meta-analysis[10] summary statistics (7771 variants).

[c]All variants with a p-value $<1 \times 10^{-5}$ identified from the re-calculated summary statistics without 23andMe (7840 variants).

[d]All linkage disequilibrium independent variants among all genotyped variants (140 467 variants).

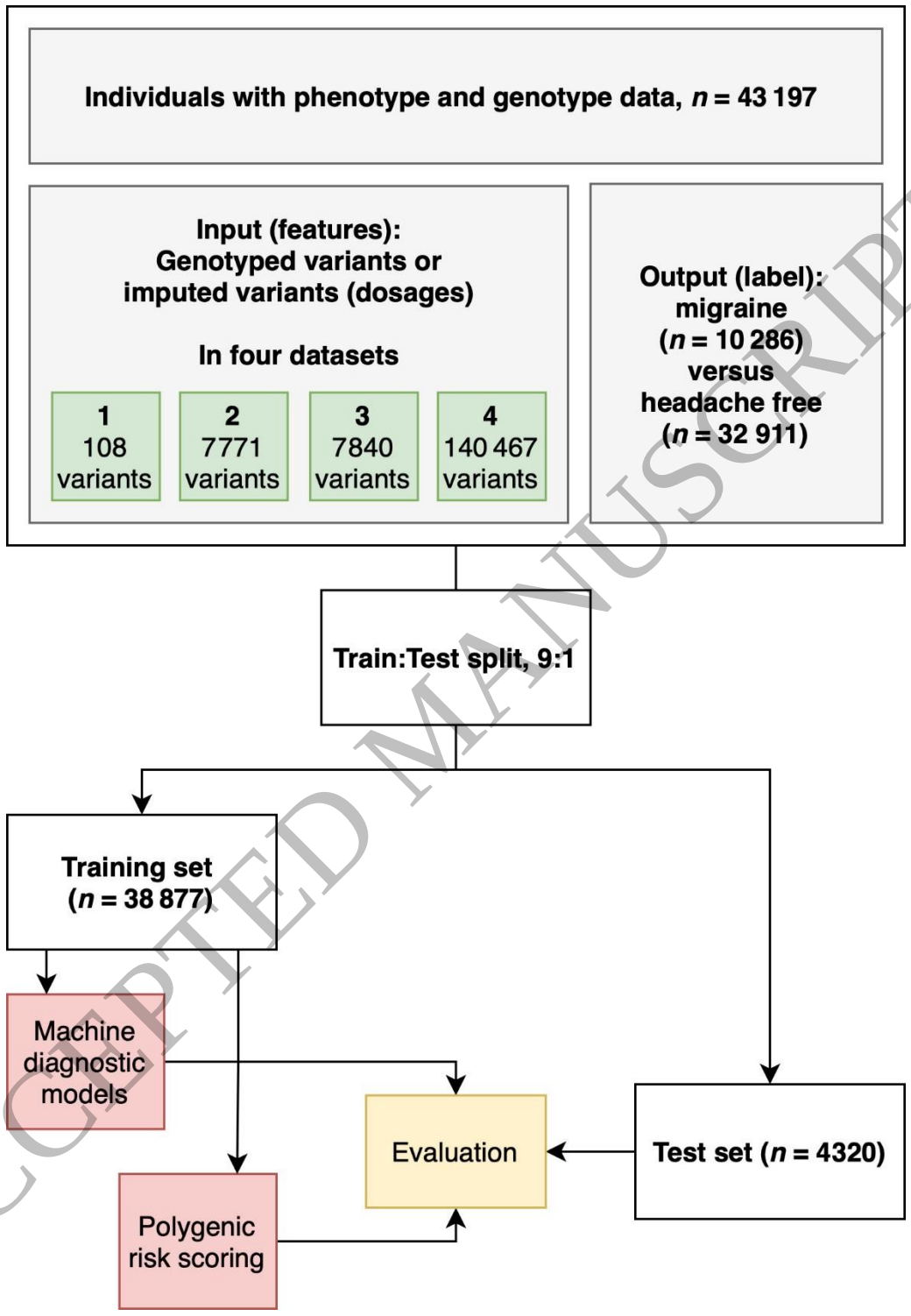

Figure 1
136x197 mm (x DPI)

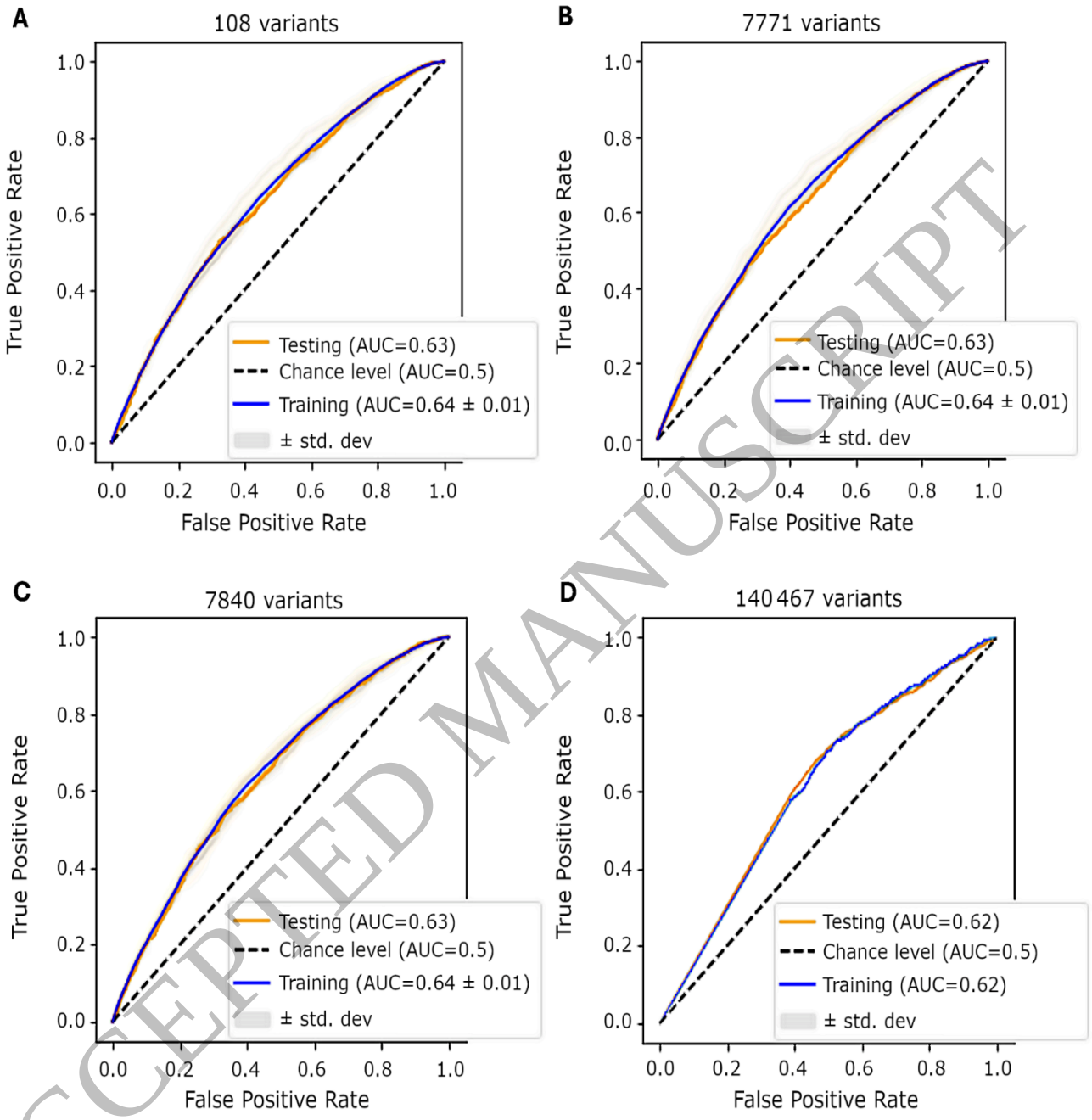

Figure 2
213x188 mm (x DPI)

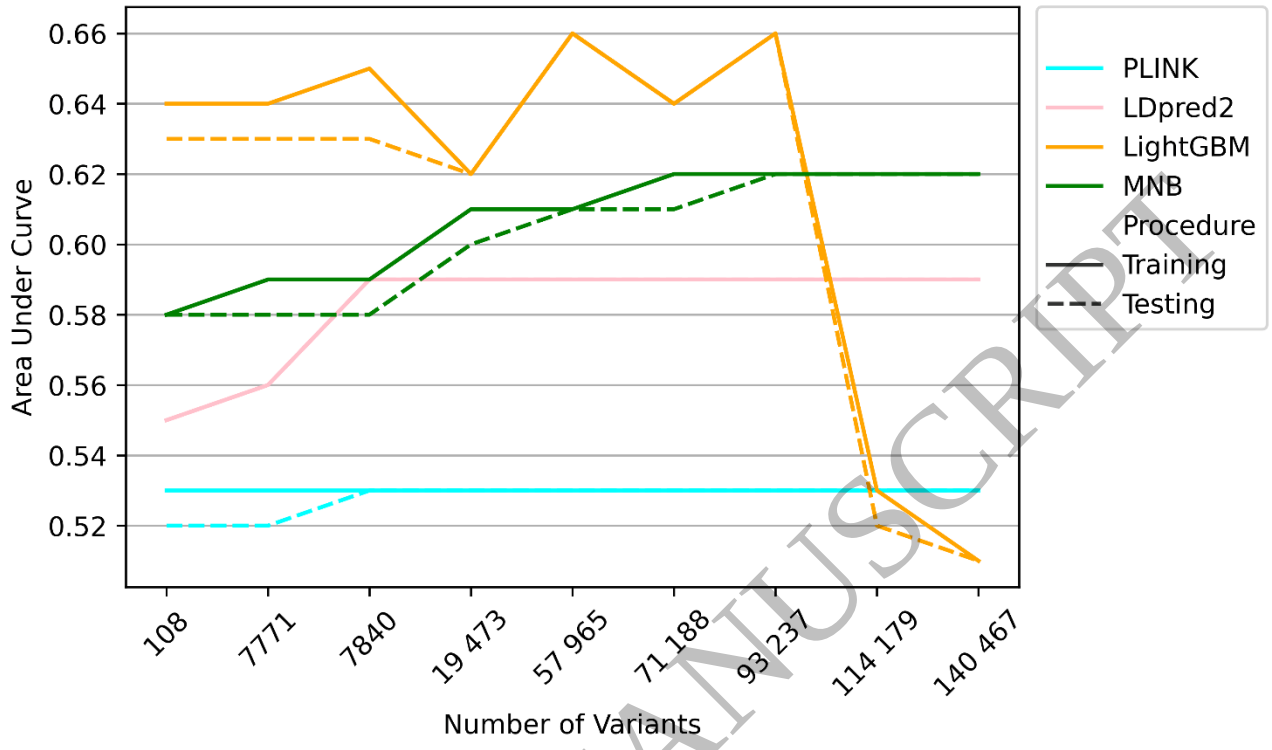

Figure 3
171x102 mm (x DPI)

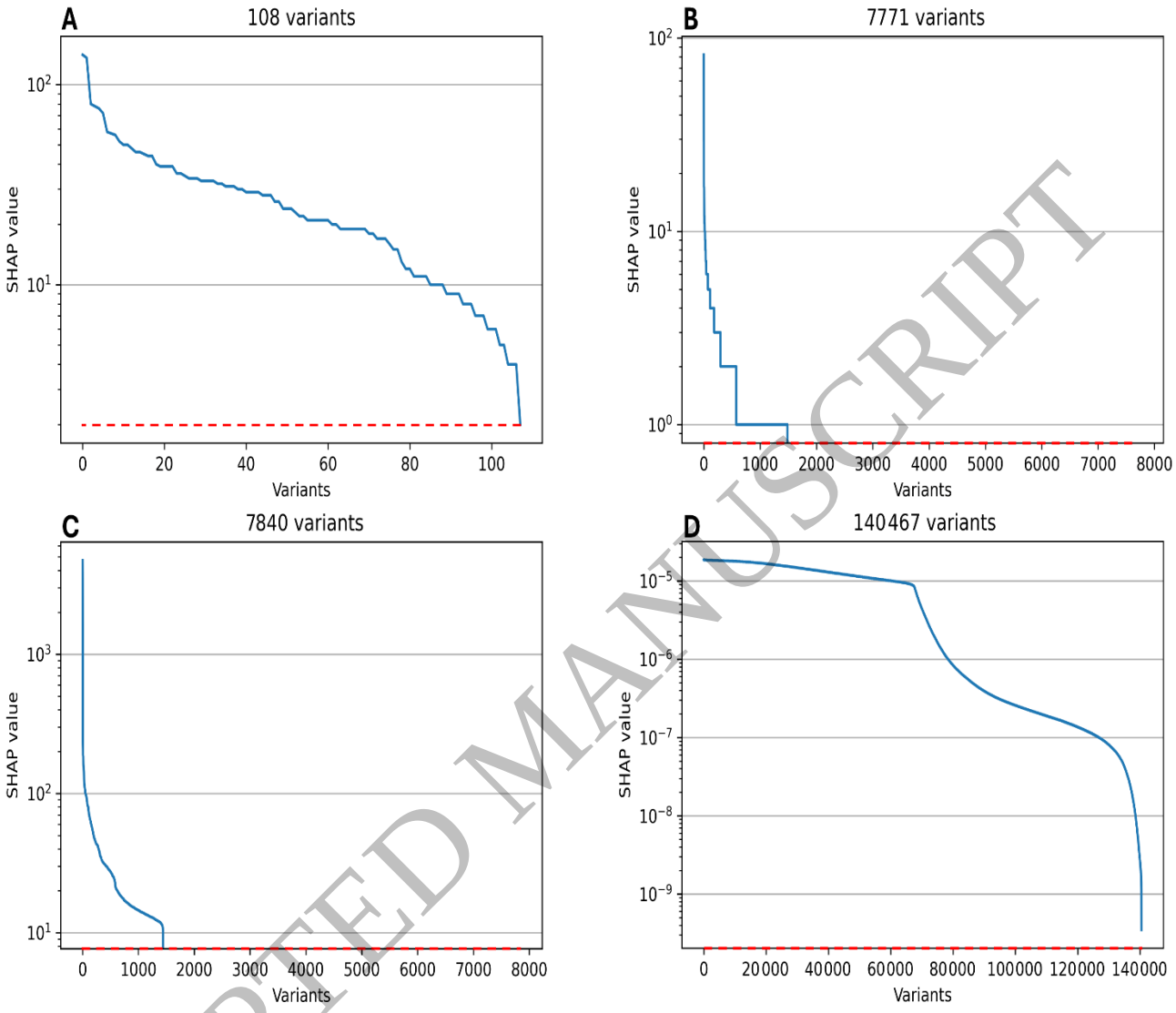

Figure 4
294x190 mm (x DPI)

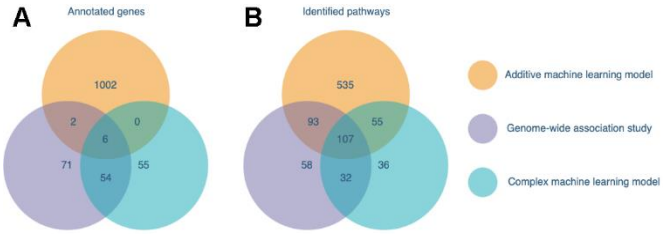

Figure 5
88x31 mm (x DPI)

TYSABRI™ (natalizumab)
SUBCUTANEOUS USE | 300mg

Prescribing Information

Efficacy made Convenient

TYSABRI SC injection with the potential to administer **AT HOME** for eligible patients*

Efficacy and safety profile comparable between TYSABRI IV and SC†[1,2]

†Comparable PK, PD, efficacy, and safety profile of SC to IV except for injection site pain.[1,2]

CLICK HERE TO DISCOVER MORE ABOUT TYSABRI SC AND THE DIFFERENCE IT MAY MAKE TO YOUR ELIGIBLE PATIENTS

Supported by

BIOGEN'S Stratify JCV™ SERVICE

A Biogen developed and funded JCV antibody index PML risk stratification service, validated and available exclusively for patients on or considering TYSABRI.

*As of April 2024, TYSABRI SC can be administered outside a clinical setting (e.g. at home) by a HCP for patients who have tolerated at least 6 doses of TYSABRI well in a clinical setting. Please refer to section 4.2 of the SmPC.[1]

TYSABRI is indicated as single DMT in adults with highly active RRMS for the following patient groups:[1,2]

- **Patients with highly active disease despite a full and adequate course of treatment with at least one DMT**
- **Patients with rapidly evolving severe RRMS defined by 2 or more disabling relapses in one year, and with 1 or more Gd+ lesions on brain MRI or a significant increase in T2 lesion load as compared to a previous recent MRI**

Very common AEs include nasopharyngitis and urinary tract infection. Please refer to the SmPC for further safety information, including the risk of the uncommon but serious AE, PML.[1,2]

Abbreviations: AE: Adverse Event; **DMT:** Disease-Modifying Therapy; **Gd+:** Gadolinium-Enhancing; **HCP:** Healthcare Professional; **IV:** Intravenous; **JCV:** John Cunningham Virus; **MRI:** Magnetic Resonance Imaging; **PD:** Pharmacodynamic; **PK:** Pharmacokinetic; **PML:** Progressive Multifocal Leukoencephalopathy; **RRMS:** Relapsing-Remitting Multiple Sclerosis; **SC:** Subcutaneous.

References: 1. TYSABRI SC (natalizumab) Summary of Product Characteristics. **2.** TYSABRI IV (natalizumab) Summary of Product Characteristics.

Adverse events should be reported. For Ireland, reporting forms and information can be found at www.hpra.ie. For the UK, reporting forms and information can be found at https://yellowcard.mhra.gov.uk/ or via the Yellow Card app available from the Apple App Store or Google Play Store. Adverse events should also be reported to Biogen Idec on MedInfoUKI@biogen.com 1800 812 719 in Ireland and 0800 008 7401 in the UK.

Biogen-261128. DOP: April 2025

Biogen