Machine learning models reveal distinct disease subgroups and improve diagnostic and prognostic accuracy for individuals with pathogenic *SCN8A* gain-of-function variants

Joshua B Hack[1], Joseph C Watkins[2], Michael F Hammer[1,3,*]

[1]BIO5 Institute, University of Arizona, Tucson AZ, USA
[2]Department of Mathematics, University of Arizona, Tucson AZ, USA
[3]Neurology Department, University of Arizona, Tucson AZ, USA

*Corresponding author's email: hammeruofa@gmail.com

Summary Statement

Application of machine learning methods to data collected by the International SCN8A Registry identifies clinically meaningful subgroups within the broad spectrum of disorders suffered by patients possessing gain-of-function variants.

ABSTRACT

Distinguishing clinical subgroups for patients suffering with diseases characterized by a wide phenotypic spectrum is essential for developing precision therapies. Patients with gain-of-function (GOF) variants in the *SCN8A* gene exhibit substantial clinical heterogeneity, viewed historically as a linear spectrum ranging from mild to severe. To test for hidden clinical subgroups, we applied two machine learning algorithms to analyze a dataset of patient features collected by the International *SCN8A* Patient Registry. We utilized two research methodologies: a supervised approach that incorporated feature severity cutoffs based on clinical conventions, and an unsupervised approach employing an entirely data-driven strategy. Both approaches found statistical support for three distinct subgroups and were validated by correlation analyses utilizing external variables. However, distinguishing features of the three subgroups within each

approach were not concordant, suggesting a more complex phenotypic landscape. The unsupervised approach yielded strong support for a model involving three partially-ordered subgroups rather than a linear spectrum. Application of these machine-learning approaches may lead to improved prognosis and clinical management of individuals with *SCN8A* GOF variants and provide insights into the underlying mechanisms of the disease.

INTRODUCTION

Rare diseases are estimated to affect approximately 1 of every 10 individuals in the human population. The low incidence coupled with the relatively recent introduction of genetic testing often means there are too few known cases upon which to establish effective protocols for diagnosis, prognosis, and treatments. In addition, the wide phenotypic spectrum often associated with many rare diseases makes it challenging to identify targeted treatment strategies (Imbrici et al., 2016). Patient registries play an indispensable role in filling gaps in the literature for rare and newly discovered diseases that have few published case reports or cohort studies. The addition of patient information coupled with advancements in computational sciences and machine learning (ML) to streamline the process of organizing, transforming, and analyzing large amounts of data hold promise for more efficient pipelines moving from gene discovery to precision medicine (Knowles et al., 2022). By using more traditional statistical methods alongside ML models, researchers can compare treatment efficacy (Cutter et al., 2019, Lee et al., 2017), predict survival probability (Mahadevan et al., 2018) and health risk factors (Taushanov et al., 2021), and improve diagnosis (Bica et al., 2021, Hack et al., 2023, Zou and Ma, 2020). These methods facilitate construction of clinical tools to describe disease and predict a patient's optimal treatment based on both genotypic and phenotypic characteristics, providing the possibility for personalized medicine.

An example of the success of the application of these methods to registry data is the case of *SCN8A*-epilepsy and related disorders (Knowles et al., 2022). Patients with pathogenic variants in the *SCN8A* gene, encoding the voltage-gated sodium channel Na$_V$1.6, exhibit substantial clinical heterogeneity, with phenotypes ranging from neurodevelopmental delays with or without seizures, to benign familial epilepsy, to a continuum of mild to severe development and epileptic encephalopathy (DEE) (Cutts et al., 2022, Hammer et al., 2023, Johannesen et al., 2022, Talwar and Hammer, 2021). The *SCN8A* gene was discovered to cause pediatric epilepsy in 2012 (Veeramah et al., 2012), and an online registry was established in 2014 that now contains an extensive dataset encompassing medical, genetic, developmental, and comorbidity

information for over 400 international cases (Andrews et al., 2023, Chung et al., 2023, Encinas et al., 2019). Given the wide phenotypic spectrum and numerous pathogenic variants in the gene, both with loss-of-function (LOF) and gain-of-function (GOF) effects, clinicians have been challenged to provide accurate diagnosis, prognosis, and a course of effective treatment. Recent advances have made it clear that the LOF and GOF subtypes differ in biological terms given the different effects LOF and GOF variants have on $Na_V1.6$ function and differential response to antiseizure medications (ASMs) (Hack et al., 2023, Johannesen et al., 2022, Liu et al., 2019, Wagnon et al., 2016). Hack et al. (2023) developed a predictive modeling approach to classify GOF and LOF variants based on clinical features present at initial diagnosis. This was an important step given that *in vitro* studies to infer channel function are not feasible in the clinic. Despite these gains, currently there are no guidelines for best practices in treating *SCN8A*-related epilepsy.

An important unanswered question is whether individuals across the phenotypic spectrum associated with variants with GOF effects can be subdivided into subgroups that differ in response to ASMs. Indeed, individuals carrying different recurrent GOF variants are known to vary in disease course, and possibly in response to ASMs (Chung et al., 2023, Hammer et al., 2023). In this study, we analyze the International *SCN8A* Registry data representing 253 cases to identify key features present early in disease progression and construct a series of predictive models to classify individuals possessing GOF variants as mild, moderate, or severe cases. Each model is assessed for accuracy via confusion matrices or mean error to test the hypothesis that the population of patients possessing GOF variants is composed of three distinct groups.

RESULTS

Loss-of-Function Exclusion

Application of the LOF Classifier (Hack et al., 2023) identified 180 of 253 individuals in the International *SCN8A* Patient Registry (Andrews et al., 2023, Encinas et al., 2019) as having a high probability of possessing a pathogenic GOF *SCN8A* variant. These 180 individuals constituted the dataset used in this study, and justification for inclusion of the variant is provided in **Table S1**. The average and standard deviation of age at onset in this subpopulation was 3.6±2.86 months and all individuals reported experiencing seizures. The average developmental quotient was 36.8±39.87 with 81% reporting developmental delay. The frequencies of seizure

types are described in Table 1. Of 115 variants, 87 were singletons, 18 were doubletons, and 10 were reported in at least three individuals. 100 variants (119 individuals) were inferred to be GOF (Hack et al., 2023) and 15 of these variants (61 individuals) were known to be GOF based on electrophysiological data.

Cluster Assignment

Unsupervised clustering analysis using partitioned around medoid (PAM) clustering showed the optimal number of clusters in this dataset was three. We denote these clusters U3, U2, and U1. This finding is consistent with the common understanding that mild, moderate, and severe groups exist in the *SCN8A* GOF population, as shown in previous work (Hack et al., 2023) (**Figure 1A**). Testing of the optimal number of clusters using kernel density estimation further supported this finding (**Figure 1B**). The average age at onset and developmental quotient (DQ; see methods for definition) were used to determine an order to the assigned clusters, and it was found that there was an increase in age at onset from Cluster U3 to Cluster U1 (**Table 1**). This suggested Cluster U3 correlated to a severe phenotype, Cluster U2 to a moderate phenotype, and Cluster U1 to a mild phenotype. Notably, developmental quotient did not follow this trend and was 26.6±32.6 and 36.5±33.1 in Clusters U3 and U1, respectively, while being higher in Cluster U2 at 65.6±47.3. (A quotient of 100 is considered neurotypical) The distribution of individuals in each category showed a strong imbalance favoring Cluster U3 (**Figure 1C**). Three clusters were also assigned as part of a *Supervised Approach* by using clinical severity cutoffs. These clusters were considered severe, moderate, and mild DEE and we denote these clusters S3, S2, and S1, respectively. Similar to the *Unsupervised Approach*, there was a strong imbalance in cluster size favoring S3 (**Figure 1D**).

Dimension Reduction

Principal component (PC) analysis on the data shows that 66.6% of the variance in the dataset was explained using the first three dimensions (**Figure 2A**). A 2-dimensional PC plot showed developmental quotient and age at onset having the greatest contribution to the variance in the dataset (**Figure 2B**). It was apparent that there was a significant overlap in the groups assigned by the *Unsupervised Approach*. However, utilizing a 3-dimensional PC plot showed increased differentiation between groups (**Figure S1**). While there were distinct centroids for each group, there was high variability within Cluster U2. The Cluster U3 group appeared to be the most uniform, with the majority being located to the left of the y-axis with its centroid close to the x-

axis. Across the three groups, there was a general trend of increasing age at onset from Cluster U3 to U1.

The *Supervised Approach* showed distinction between the three groups in 2 dimensions (**Figure 2C**). Individuals in Cluster S3 were centered to the left of the y-axis and above the x-axis. Individuals in the Cluster S2 or Cluster S1 group were centered progressively more positively along the x-axis with a clearer division between the two phenotypic groups than seen in the *Unsupervised Approach*. Most individuals in Cluster S1 were also above the x-axis, suggesting better development and lower rates of bilateral tonic clonic (BTC) and convulsive seizures at initial presentation. The distinction between each group was more visible in a 3-dimensional PC plot (**Figure S2**).

Penalized Ordinal Logistic Regression Model

A penalized ordinal logistic regression model (p-ORM) was constructed for both the *Unsupervised* and *Supervised Approaches*. Both approaches utilized *k*-fold cross-validation where *k*=5. In the *Unsupervised p-ORM*, the misclassification error across five iterations was 0.105±0.04. Tuning of the model resulted in seizure freedom, DQ and developmental delay, age at onset, infantile spasms, convulsive, myoclonic, absence, focal, and tonic seizures being selected (**Table 2**). Following penalization, the optimal tuning parameter (best lambda index) was λ=18 and a confusion matrix was constructed using this parameter, resulting in an error of 0.089 (**Table 3**). The *Supervised p-ORM* resulted in a misclassification error of 0.144±0.04. Features that were selected by tuning the model included age at onset, seizure freedom, DQ, tonic, convulsive, and BTC seizures (Table 2). Penalization resulted in λ=14 and a confusion matrix with an error of 0.139 (**Table 3**).

Stacked Meta-Learner

Both datasets were analyzed using a stacked meta-learner (Stacked) three times. The dataset was first sampled in five separate ways before a random forest model was trained and tested using *k*-fold cross validation where *k*=4 using the methods described in (Desprez et al., 2022). The predicted outcome classification from each sampling method were utilized as training features in an ordinal logistic regression model while the probabilities were used in a linear regression model as a meta-learner. In the *Unsupervised Stacked Model*, the mean absolute error (MAE) was 0.06±0.01 and 0.04±0.03 for, respectively, the outcome classification and probability from the random forest model (**Table 4**). The root-mean-square error (RMSE) was 0.16±0.01 and 0.15±0.07 and the percentage of correct classification (PCC) 94.8±0.8 per cent

and 95.2±2.70 per cent for classification and probability, respectively. In the *Supervised Stacked Model*, performance resulted in MAEs of 0.37±0.09 and 0.35±0.07, RMSEs of 0.54±0.07 and 0.52±0.08, and PCCs of 81.0±4.58 per cent and 81.0±1.00 per cent for classification and probability, respectively (**Table 4**).

External Validation

To determine if the *Unsupervised Stacked* and *Supervised p-ORM* have biological and clinical relevance, each model was validated using external data (Johannesen et al., 2022). Biological relevance was tested using variants that have undergone electrophysiological studies and have an associated score based on 6 distinctive features. Linear regressions using values from the *Unsupervised Stacked* model showed a significant correlation against a test of no correlation between Cluster U3 and electrophysiological score and the Combined score and electrophysiological score (p-value=0.016 and 0.022, respectively, and adjusted R-squared= 0.094 and 0.084, respectively) (**Figure 3A-D**). Additionally, the probability of being Cluster U1 showed a nearly significant relationship with electrophysiological score (p-value=0.054; adjusted R-squared= 0.055). R1872W has been suggested to be an outlier regarding its electrophysiological score and phenotypic severity, potentially due to its difference in peak current from R1872Q, which was not included in the calculation of electrophysiological score (Chung et al., 2023, Johannesen et al., 2022). To account for this, regression was again performed excluding R1872W. All p-values were found to decrease while R-squared values increased, with Cluster U3, Cluster U1, and Combined showing significant correlation against a test of no correlation (p-value= 0.002, 0.010, and 0.004, respectively, and adjusted R-squared= 0.207, 0.141, and 0.181, respectively).

In the *Supervised p-ORM*, linear regressions of all four severity scores show significant relationships with electrophysiological score with p-values of 2.84×10^{-5}, 1.95×10^{-5}, 0.022, and 2.48×10^{-4} for Cluster S3, S2, S1, and Combined, respectively, and adjusted R-squared of 0.266, 0.276, 0.077, and 0.208, respectively (**Figure 3E-H**). Further validation of the *Supervised p-ORM* was conducted using a dataset of individuals with clinical decisions made regarding their severity (Johannesen et al., 2022). The *Supervised p-ORM* was used to predict on this dataset, using the same features. Prediction resulted in a mean error of 0.209.

Discussion

The application of machine learning techniques resulted in statistical support for three distinct patient subgroups within the wide phenotypic spectrum associated with *SCN8A* GOF variants. Utilizing features with severity cutoffs that were determined by clinical conventions, the *Supervised Approach* categorized individuals into three groups that supported the historical view of Mild, Moderate, and Severe DEE subgroups. The *Unsupervised Approach* confirmed the presence of three disease GOF subgroups using a strictly data-driven strategy (i.e., without the bias of prior clinical interpretation). However, the distinguishing features of the three subgroups were not concordant between the two approaches. A key finding of the Unsupervised Approach is the clinical separation of epileptic and developmental encephalopathies. In the following sections we discuss the pros and cons of each approach and the significance of the finding of distinct disease subgroups.

Both the *Unsupervised* and *Supervised Approaches* were modeled using penalized ordinal logistic regression to identify features that were critical to distinguishing the three clusters and again using a stacked meta-learner to account for sample size imbalance within groups. In the *Unsupervised Approach* it was found that both modeling methods performed well in classifying individuals, with the average error being 0.11 and 0.04 for the p-ORM and stacked models, respectively (**Tables 3-4**). The primary concern for both the *Unsupervised Stacked* and *Unsupervised p-ORM* is the potential for over-reliance on seizure freedom as a predictive feature. While the p-ORM shows a strong impact from seizure freedom (**Table 2**), the effect is limited in the *Unsupervised Stacked*. This, taken together with the improvement in performance metrics, makes the stacked meta-learner the preferred choice.

In the *Supervised Approach*, only the p-ORM performed well enough to be considered viable in further analysis. This model recorded an average error of 0.144 across the five folds in cross validation and a final error of 0.139 after penalization (**Table 3**) compared to the average MAE of 0.35 observed in the Stacked model (**Table 4**). In this approach, the complexity of the Stacked meta-learner may lead to a decrease in overall performance. As such, the *Supervised Approach* is best modeled using the penalized ORM. This model selected age at seizure onset as its primary feature contributing to predicted outcome, as expected from the original classification guidelines (**Table 2**).

Both the *Unsupervised* and *Supervised Approaches* were tested for biological relevance following the success of modeling. In the case of the *Unsupervised Stacked* model, the predicted probabilities of an individual being in Cluster U3 and U1, as well as a combined probability score across all three clusters, were shown to have a significant linear relationship with electrophysiological score calculated for highly recurrent *SCN8A* variants (**Fig. 3**). Similarly, the *Supervised p-ORM* showed significant relationships between all categories and variant-specific electrophysiological score (**Fig. 3**). The variant electrophysiological score has previously been shown to correlate with clinical severity (Chung et al., 2023, Johannesen et al., 2022). While this is a simplified surrogate parameter, the congruence between these electrophysiological scores and predicted outcome lends additional support to the validity of these classifications.

Additional relevance testing was conducted on the *Supervised p-ORM* using an externally curated dataset (Johannesen et al., 2022) to test the performance of the model. In this case, the error of the predictive model increased from 0.139 to 0.209. This decrease in performance is expected, as the dataset used for external validation collected data in a different format than the data used in this study. For example, developmental quotient was not available for individuals in the external dataset and was collected as categorical values of intellectual and developmental disorder, rather than a continuous variable. Individuals were given generalized values in instances of data collection discrepancies, which can lead to underfitting in a ML model.

Discordance in Clusters

The results of the *Unsupervised Approach* to identifying and assigning clusters provide a novel method for classifying individuals. In the case of the *Supervised Approach*, clusters correspond to Mild DEE, Moderate DEE, and Severe DEE as described (Hammer et al., 2023). As expected, DQ follows a trend of decreasing from Cluster S1 to S3. Seizure-free individuals are present in all three clusters. Strikingly, the *Unsupervised Approach* is not concordant with these results. In Cluster U1, DQ is lower in both Cluster U2 and Cluster U3. This contrasts with age at onset is higher than Cluster U2 but lower in Cluster U3. Moreover, *all* individuals experience seizure freedom in Cluster U1. Cluster U2 is characterized by no to moderate developmental delay with an earlier average age at onset than Cluster U1. Cluster U3 shows severe developmental delay and exceedingly early age at onset with no individuals experiencing seizure freedom. Those who attained prolonged seizure freedom likely benefitted from effective

treatment with ASMs. It should be noted that SeLFIE (self-limited(familial) infantile epilepsy) is an additional phenotype that was not observed in the Registry.

The marked non-concordance between the two approaches challenges the conventional understanding of the *SCN8A* GOF population as an ordered spectrum from mild to severe DEE. Taken as a whole, Cluster U1 experiences the latest average age at seizure onset, while also having few relative instances of more aggressive seizure etiologies such as myoclonic seizures. Also, this group consistently achieves seizure freedom. These characteristics tend to align with the expectation for a mild DEE, but the severe developmental delay indicated by a low average DQ conflicts with this understanding under the presently used classification of *SCN8A* epilepsy. However, Cluster U2 is characterized by a moderate average age at onset with a low seizure freedom rate. These seizure characteristics are consistent with a moderate DEE, but fewer than a quarter (10/43 individuals) report developmental delay within this group, once again being at odds with the current moderate DEE classification. Cluster U3 is the only group largely concordant with its expected counterpart S3 in the *Supervised Approach* in both DQ and age at onset. Notably, this group reports zero incidents of seizure freedom compared to the 11 incidents of seizure freedom reported in Cluster S3 (**Table 1**).

The notion that individuals with *SCN8A* GOF do not neatly fit a linear spectrum of increasingly severe DEE is not entirely unexpected. The distinction between mild and moderate DEE has been a consistent challenge for clinicians, leading to these two categories often being collapsed (Hammer et al., 2023). As such, we hypothesize that the three clusters identified using the *Unsupervised Approach* are a partially ordered subgroup set where Cluster U1 is primarily a developmental encephalopathy (DE), Cluster U2 is primarily an epileptic encephalopathy (EE), and Cluster U3 is a DEE. Given this hypothesis, it is predicted that individuals in the DE population typically would experience developmental delay prior to seizure onset, while those in EE would experience seizure onset prior to developmental delay. Those in DEE category would be expected to have clinical diagnoses of both seizure onset and developmental delay nearly simultaneously within the first few months of life.

Our results are in line with previous studies demonstrating correlations between specific pathogenic variants and severity outcome (Chung et al., 2023). For example, the most debilitating variants- p.R850Q and p.R1872W- were classified as DEE 90.5% of the time (19/21), while variants that are associated with milder clinical outcomes such as p.G1475R and

p.N1877S were classified as either DE or EE 83.3% of the time (10/12). This provides a second line of support for a partially-ordered model in which DEE is the most severe subgroup, with similar degrees of severity for DE and EE. Thus, a strictly ordinal model is unlikely to be the most appropriate for classifying individuals with GOF variants. As more phenotypic data become available from individuals with highly recurrent variants, our model of subgroup classification may reveal a deeper understanding of the underlying mechanisms of the disease. In addition, the mechanistic properties of different GOF variants as inferred from *in vitro* electrophysiological studies may correlate with particular disease subgroups (Chung et al., 2023).

Implications for Treatment

This study provides a critical insight for clinicians to use in assessing a patient's prognosis. The two approaches utilized here result in models that perform with high accuracy and identify features that are important to understanding possible patient outcomes. These features are present early in the disease progression including age at seizure onset, developmental quotient, developmental delay, initial seizure type, and if seizures have been controlled for at least six months. A strength of this model is that all features except prolonged seizure freedom can be assessed accurately in the first clinical visit following genetic testing. As such, it is believed that this model is particularly useful for improving prognosis. Not only do these models perform well based on features revealed early in life, but they have also been shown to have biological relevance as determined by regression on outcome probability versus electrophysiological score.

Given the two different approaches to classifying individuals, one model from each approach is proposed in aiding clinical decisions. In the case of the *Supervised Approach*, the penalized ordinal regression model performs with high accuracy and considers age at onset, DQ, and three seizure types as the most highly predictive features (**Table 2**). The low error and data requirement for this model makes it particularly well suited to clinical settings, especially compared to the *Supervised Stacked* model, which has a high error (**Table 4**). One important consideration for this model is the unbalanced cluster assignments that may lack power to detect some phenotype-phenotype associations. This is overcome by using the data-driven *Unsupervised Approach*. Not only is the bias limited in the *Unsupervised Approach*, both models using these clusters perform with lower error than the *Supervised p-ORM* (**Tables 3-4**). A shortfall of the *Unsupervised p-ORM* is its heavy reliance on seizure freedom (**Table 2**). Because of this, the *Unsupervised Stacked* model should be used if relying on this approach.

Further support of the hypothesis that the *SCN8A* GOF population can be categorized as DE, EE, or DEE will favor the use of the *Unsupervised Stacked* model over the *Supervised p-ORM*. This would have implications for effective treatment and would help to inform clinicians and caretakers of realistic treatment goals. In the case of DE, effective treatment may result in attainment of prolonged seizure freedom without notable improvements in developmental progression. On the other hand, despite seizures being more difficult to control for individuals with EE some improvements in acquisition of developmental milestones may come with reduction in seizure burden. If an affected individual is classified as DEE, effective treatment may reduce seizure burden without achievement of seizure freedom or improvements in acquisition of developmental skills. As we acquire additional patient reports in support of this classification system, then the expectation is that this classification will become more standard among clinicians and thus become the supervised model for the next refinement in *SCN8A* classification.

Caveats and Significance

While both models are effective in accurately classifying individuals into their respective groups, these models are limited to the datasets available to them at training. As these models are further developed as more longitudinal data become available, it is expected that the models will continue to be refined and improve in accuracy. Currently, their utility is mainly as a tool to aid in clinical decisions. Despite these limitations, we believe that these models can be used in combination with the LOF Classifier (Hack et al., 2023) to provide rapid predictions of variant effect and expected patient subgroup. The results from the *Unsupervised Approach* also lay the foundation for a potential shift in the understanding of the phenotypic landscape of *SCN8A* Syndrome, with possible implications for disease mechanisms. The results from this work may help with developing more attuned machine learning models to better characterize the entire phenotypic spectrum, including both LOF and GOF variants with a variety of biophysical effects on the $Na_V1.6$ channel. This will aid clinicians in making treatment decisions and provide more realistic expectations for caregivers of individuals with *SCN8A*-related disorders.

The construction of this predictive model provides the opportunity for further development of machine learning tools to assist in clinical settings. However, the field of functional prediction is still emerging and while these models show promise in a research capacity, they should not be used to make decisions on clinical management. Before widespread use of these models

becomes a viable option in clinical settings, trials must be conducted to validate the results and test the robustness of these models. To emphasize, these models are not suitable replacements for clinical guidelines on managing *SCN8A*-related disorders, but rather tools to aid in research and may eventually be validated for clinical use. Clinicians who are interested in using these models should be aware of their limitations and use the results as only a part of their therapeutic strategies. Ultimately, these models are tools that may develop to higher magnitudes of utility but presently should be used with caution and awareness of the limitations and risks outside of a research setting.

Methods

Data Collection

Data for this study were collected by the International *SCN8A* Patient Registry (Andrews et al., 2023, Encinas et al., 2019). 253 responses during the period from January 2017 through December 2021 from consenting participants were considered for this study. Using this dataset, features were selected that are present and easily assessed in clinical settings at early stages of the disease: genetic variant, age at onset, seizure history, initial type(s) of seizure(s), current type(s) of seizure(s), developmental quotient (DQ), presence of developmental delay, and whether the patient had experienced a period of at least 6 months of seizure freedom. DQ was calculated by dividing developmental age by patient's age at the time of their registry submission multiplied by 100. Developmental age was calculated using 25 skills from the Denver II Developmental Test that are queried in the Registry. This quotient uses the 90th percentile for the age that a given skill is acquired by a neurotypical child. Initial and current seizure types were expanded from a list variable to a single binary variable for each seizure type.

Inclusion Criteria

Individuals were included in this study only if they possessed *SCN8A* variants with known or inferred gain-of-function (GOF) properties and were classified as pathogenic or likely pathogenic according to genetic reports, the ClinVar database, or according to American College of Medical Genetics (ACMG) guidelines (Richards et al., 2015). Variants known to have GOF properties from previous electrophysiological studies performed in heterologous expression systems were also included. Additional variants were included by utilizing the LOF Classifier (Hack et al., 2023) to categorize variants: those with a probability of loss of function (LOF) [prob(LOF)] <0.3

were considered GOF, while those with a prob(LOF) > 0.3 were considered intermediate or true LOF variants and excluded from the study. Alternative methods exist for classifying variants as GOF or LOF (Bosselmann et al., 2023, Brunklaus et al., 2022, Heyne et al., 2020). After reviewing these alternatives, it was determined that the LOF classifier from Hack et al., 2023 was sufficient as justification for inclusion of variants in a clinically focused study. The alternative methods were determined to be better suited for evolutionary and comparative studies of sodium channels or in cases which a disease has limited available clinical data. Additionally, individuals with a pathogenic or likely pathogenic (but not benign or likely benign) variant at other loci associated with epilepsy were excluded from the study. The resulting dataset consisted of 180 individuals.

Cluster Assignment and Data Visualization

Unsupervised Approach. Toward the goal of limiting clinician bias and discovering new phenotype-phenotype links, unsupervised cluster analysis and assignment was completed. Individuals were assigned clusters by using a dissimilarity matrix calculation with a generalized Gower formula resulting in an optimal number of 3 clusters, named U1, U2, and U3. This was determined using partitioned around medoid (PAM) clustering and further testing of these clusters using hierarchical clustering to validate these assignments. A range of 1-5 was tested for the optimal number of clusters and further validation of cluster assignments was conducted using density-based spatial clustering of applications with noise (DBSCAN). Analysis of each cluster's features suggested that Cluster U1 associated with milder phenotypic outcome while Cluster U3 associated with more severe outcomes, with Cluster U2 being intermediate. These clusters were used as the response feature in each predictive model as an *Unsupervised Approach.*

Supervised Approach. To ensure clinical relevance, assessment of patient health severity was determined by a clinical specialist. These group assignments were determined by age at onset, DQ, and a combination of initial seizure type and number of seizure types at initial presentation. The severity groups of Mild (Cluster S1), Moderate (Cluster S2), and Severe (Cluster S3) corresponded generally to the Mild, Moderate, and Severe DEE categories described in (Hammer et al., 2023), respectively. There were no instances of the classification described as self-limited familial infantile epilepsy (SeLFIE) (Hammer et al., 2023). These severity groups were used as the response feature in each predictive model as a *Supervised Approach.*

Predictive Modeling

A series of predictive models were constructed using clusters assigned by both the *Unsupervised* and *Supervised Approaches.* In both approaches, penalized ordinal logistic regression (p-ORM) and a stacked meta-learner utilizing a random forest classifier, an ordinal logistic regression classifier, and a linear regression model as described in (Desprez et al., 2022)(Stacked) were used to model the dataset. Each model used Clusters as the response feature. In the case of the p-ORM and Stacked models, predictor features were age at onset, DQ, seizure freedom, initial seizure type(s), and developmental delay. The Stacked model accounts for imbalance between groups by constructing 5 random forest models using conventional sampling methods for contribution to predictive performance. These methods include no sampling, oversampling, undersampling, over/undersampling, and synthetic minority oversampling technique (SMOTE) sampling. The random forest probability and classification outputs of the selected sampling methods are then used as inputs in a linear and ordinal regression model, respectively, for a final classification.

For the p-ORM, *k*-fold cross validation was used with five folds to estimate an average error across the entire model. The optimal penalization parameter, λ, was calculated for both approaches. Similarly, the Stacked model utilized *k*-fold cross validation by splitting the training set into three folds. Each model was trained and tested on the holdout data and mean absolute error (MAE), root-mean-square deviation (RMSE), and percentages of correct classification (PCC) were used to evaluate the performance of each model. To ensure that the Stacked model's performance was not a result of favorable training holdouts, each Stacked model was run with three different seeds and the average MAE and RMSE from each run was calculated for a final performance value.

External Validation

To determine biological and clinical relevance, the Stacked model using the *Unsupervised Approach* and the p-ORM using the *Supervised Approach* were validated. Both models were validated for biological relevance by taking a subset (*n*=51) of the population possessing variants that have associated electrophysiological scores based on 6 distinctive features: half activation voltages of activation and inactivation curves, slopes of the activation and inactivation curves, maximum conductance, and persistent current reported in (Johannesen et al., 2022). Recent work has shown these electrophysiological scores to be correlated with clinical severity (Chung et al., 2023). A dataset was constructed with these individuals that included the

predicted probability of cluster assignment for any sampled ~~that~~ patient for the Stacked model in the case of the *Unsupervised Approach* and the p-ORM in the case of the *Supervised Approach.* Linear regression was performed on each of these features using the electrophysiological score associated with each variant as the response feature. For the explanatory feature, an expected severity score for each patient was calculated by multiplying the probabilities of the patient being Cluster U1/S1, Cluster U2/S2, and Cluster U3/S3 by one, two, and three, respectively.

Clinical validation was deemed necessary for the *Supervised Approach* as the cluster assignments for this approach follow clinical conventions, while the *Unsupervised Approach* follows an alternative decision-making process. A new dataset (*n*=91) was constructed using individuals reported in Johannesen et al. (Johannesen et al., 2022) with features that correspond to the prediction features of the p-ORM. Individuals from this dataset were Cluster S1 if their phenotype was reported as benign familial infantile epilepsy (BFIE) or infantile epilepsy (IE), Cluster S2 if reported as DEE with Mild-Moderate ID, and Cluster S3 if reported as DEE with Severe ID. To convert the categorical variable of ID into a continuous variable as DQ, severe ID was set to have a DQ of 40, moderate ID a DQ of 67.5, mild ID a DQ of 90, and normal a DQ of 100. Additionally, seizure freedom was inferred if the patient was classified as BFIE or IE. Using the p-ORM trained in the *Supervised Approach,* these 91 individuals were predicted as Cluster S1, Cluster S2, or Cluster S3 and a confusion matrix was constructed to assess performance with error as the primary assessment feature.

Data and Code Availability

All coding was completed in RStudio v4.2.2 and de-identified datasets and markdowns for dataset construction, expansion, analysis, and model validation are archived on Zenodo (DOI: 10.5281/zenodo.8336484) (Bischl et al., 2016, Bischl et al., 2018, Bischl et al., 2020, Fox and Hong, 2010, Lunardon et al., 2014 , OpenML, 2020, Probst et al., 2017 , Venables and Ripley, 2002 , Wright and Ziegler, 2017)

Ethics

This study was approved by the University of Arizona Institutional Review Board (#1603487278) and all caregivers of individuals with *SCN8A* consented to participate prior to providing their filling out the International *SCN8A* Patient Registry.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest.

References

ANDREWS, J. G., GALINDO, M. K., HACK, J. B., WATKINS, J. C., CONECKER, G. A., HAMMER, M. F. & CONSORTIUM, S. A. R. 2023. The International SCN8A Patient Registry: A Scientific Resource to Advance the Understanding and Treatment of a Rare Pediatric Neurodevelopmental Syndrome. *J Registry Manag,* 50**,** 4-10.

BICA, I., ALAA, A. M., LAMBERT, C. & VAN DER SCHAAR, M. 2021. From Real-World Patient Data to Individualized Treatment Effects Using Machine Learning: Current and Future Methods to Address Underlying Challenges. *Clin Pharmacol Ther,* 109**,** 87-100.

BISCHL, B., KÜHN, T. & SZEPANNEK, G. 2016. On Class Imbalance Correction for Classification Algorithms *In:* LÜBBECKE, M., KOSTER, A., LETMATHE, P., MADLENER, R., PEIS, B. & WALTHER, G. (eds.) *Credit Scoring.* Cham: Springer International Publishing.

BISCHL, B., LANG, M., KOTTHOF, L., SCHIFFNE, J., RICHTER, J., STUDERUS, E., CASALICCHIO, G. & JONES, Z. M. 2020. mlr: Machine Learning in R.

BISCHL, B., RICHTER, J., BOSSEK, J., HORN, D., THOMAS, J. & LANG, M. 2018. mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions.

BOSSELMANN, C. M., HEDRICH, U. B. S., LERCHE, H. & PFEIFER, N. 2023. Predicting functional effects of ion channel variants using new phenotypic machine learning methods. *PLoS Comput Biol,* 19**,** e1010959.

BRUNKLAUS, A., FENG, T., BRUNGER, T., PEREZ-PALMA, E., HEYNE, H., MATTHEWS, E., SEMSARIAN, C., SYMONDS, J. D., ZUBERI, S. M., LAL, D. & SCHORGE, S. 2022. Gene variant effects across sodium channelopathies predict function and guide precision therapy. *Brain,* 145**,** 4275-4286.

CHUNG, K. M., HACK, J., ANDREWS, J., GALINDO-KELLY, M., SCHREIBER, J., WATKINS, J. & HAMMER, M. F. 2023. Clinical severity is correlated with age at seizure onset and biophysical properties of recurrent gain of function variants associated with SCN8A-related epilepsy. *Epilepsia,* 64**,** 3365-3376.

CUTTER, G., XIN, H., ABAN, I., BURNS, T. M., ALLMAN, P. H., FARZANEH-FAR, R., DUDA, P. W. & KAMINSKI, H. J. 2019. Cross-sectional analysis of the Myasthenia Gravis Patient Registry: Disability and treatment. *Muscle Nerve,* 60**,** 707-715.

CUTTS, A., SAVOIE, H., HAMMER, M. F., SCHREIBER, J., GRAYSON, C., LUZON, C., BUTTERFIELD, N., PIMSTONE, S. N., AYCARDI, E., HARDEN, C., YONAN, C., JEN, E., NGUYEN, T., CARMACK, T. & HAUBENBERGER, D. 2022. Clinical characteristics and treatment experience of individuals with SCN8A developmental and epileptic encephalopathy (SCN8A-DEE): Findings from an online caregiver survey. *Seizure,* 97**,** 50-57.

DESPREZ, M., ZAWADA, K. & RAMP, D. 2022. Overcoming the ordinal imbalanced data problem by combining data processing and stacked generalizations. . *Mach. Learn. Appl. ,* 7**,** 1-9.

ENCINAS, A. C., MOORE, I. K. M., WATKINS, J. C. & HAMMER, M. F. 2019. Influence of age at seizure onset on the acquisition of neurodevelopmental skills in an SCN8A cohort. *Epilepsia,* 60**,** 1711-1720.

FOX, J. & HONG, J. 2010. Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package. . *J. Stat. Softw.,* 32**,** 1–24.

HACK, J. B., HORNING, K., JUROSKE SHORT, D. M., SCHREIBER, J. M., WATKINS, J. C. & HAMMER, M. F. 2023. Distinguishing Loss- and Gain-of-Function SCN8A Variants Using a Random Forest Classification Model Trained on Clinical Features. *Neurol Genet,* 9**,** 1-10.

HAMMER, M. F., XIA, M. & SCHREIBER, J. M. 2023. SCN8A-Related Epilepsy and/or Neurodevelopmental Disorders. *In:* ADAM, M. P., MIRZAA, G. M., PAGON, R. A., WALLACE, S. E., BEAN, L. J. H., GRIPP, K. W. & AMEMIYA, A. (eds.) *GeneReviews((R)).* Seattle (WA).

HEYNE, H. O., BAEZ-NIETO, D., IQBAL, S., PALMER, D. S., BRUNKLAUS, A., MAY, P., EPI, C., JOHANNESEN, K. M., LAUXMANN, S., LEMKE, J. R., MOLLER, R. S., PEREZ-PALMA, E., SCHOLL, U. I., SYRBE, S., LERCHE, H., LAL, D., CAMPBELL, A. J., WANG, H. R., PAN, J. & DALY, M. J. 2020. Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci Transl Med,* 12.

IMBRICI, P., LIANTONIO, A., CAMERINO, G. M., DE BELLIS, M., CAMERINO, C., MELE, A., GIUSTINO, A., PIERNO, S., DE LUCA, A., TRICARICO, D., DESAPHY, J. F. & CONTE, D. 2016. Therapeutic Approaches to Genetic Ion Channelopathies and Perspectives in Drug Discovery. *Front Pharmacol,* 7**,** 121.

JOHANNESEN, K. M., GARDELLA, E., AHRING, P. K. & MOLLER, R. S. 2022. De novo SCN3A missense variant associated with self-limiting generalized epilepsy with fever sensitivity. *Eur J Med Genet,* 65**,** 104577.

KNOWLES, J. K., HELBIG, I., METCALF, C. S., LUBBERS, L. S., ISOM, L. L., DEMAREST, S., GOLDBERG, E. M., GEORGE, A. L., JR., LERCHE, H., WECKHUYSEN, S., WHITTEMORE, V., BERKOVIC, S. F. & LOWENSTEIN, D. H. 2022. Precision medicine for genetic epilepsy on the horizon: Recent advances, present challenges, and suggestions for continued progress. *Epilepsia,* 63**,** 2461-2475.

LEE, J. M., RHEE, T. M., HAHN, J. Y., HWANG, D., PARK, J., PARK, K. W., KIM, H. L., KIM, S. H., CHAE, I. H., DOH, J. H., JEON, K. H., CHOI, Y. J., PARK, J. S., CHOI, S. H., GWON, H. C., KOO, B. K., ALFONSO, F. & KIM, H. S. 2017. Comparison of outcomes after treatment of in-stent restenosis using newer generation drug-eluting stents versus drug-eluting balloon: Patient-level pooled analysis of Korean Multicenter in-Stent Restenosis Registry. *Int J Cardiol,* 230**,** 181-190.

LIU, Y., SCHUBERT, J., SONNENBERG, L., HELBIG, K. L., HOEI-HANSEN, C. E., KOKO, M., RANNAP, M., LAUXMANN, S., HUQ, M., SCHNEIDER, M. C., JOHANNESEN, K. M., KURLEMANN, G., GARDELLA, E., BECKER, F., WEBER, Y. G., BENDA, J., MOLLER, R. S. & LERCHE, H. 2019. Neuronal mechanisms of mutations in SCN8A causing epilepsy or intellectual disability. *Brain,* 142**,** 376-390.

LUNARDON, N., MENARDI, G. & TORELLI, N. 2014 ROSE: a Package for Binary Imbalanced Learning. . *R J. 6, 79.*

MAHADEVAN, A., BLANCK, O., LANCIANO, R., PEDDADA, A., SUNDARARAMAN, S., D'AMBROSIO, D., SHARMA, S., PERRY, D., KOLKER, J. & DAVIS, J. 2018. Stereotactic Body Radiotherapy (SBRT) for liver metastasis - clinical outcomes from the international multi-institutional RSSearch(R) Patient Registry. *Radiat Oncol,* 13**,** 26.

OPENML 2020. OpenML: An R package to connect to the machine learning platform *Cluster: Cluster Analysis Basics and Extensions.* .

PROBST, P., AU, Q., CASALICCHIO, G., STACHL, C. & BISCHL, B. 2017 Multilabel Classification with R Package mlr. R J. 9, 352.

TALWAR, D. & HAMMER, M. F. 2021. SCN8A Epilepsy, Developmental Encephalopathy, and Related Disorders. *Pediatr Neurol,* 122**,** 76-83.

TAUSHANOV, Z., VERLOO, H., WERNLI, B., DI GIOVANNI, S., VON GUNTEN, A. & PEREIRA, F. 2021. Transforming a Patient Registry Into a Customized Data Set for the Advanced Statistical Analysis of Health Risk Factors and for Medication-Related Hospitalization Research: Retrospective Hospital Patient Registry Study. *JMIR Med Inform,* 9**,** e24205.

VEERAMAH, K. R., O'BRIEN, J. E., MEISLER, M. H., CHENG, X., DIB-HAJJ, S. D., WAXMAN, S. G., TALWAR, D., GIRIRAJAN, S., EICHLER, E. E., RESTIFO, L. L., ERICKSON, R. P. & HAMMER, M. F. 2012. De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am J Hum Genet,* 90**,** 502-510.

VENABLES, W. N. & RIPLEY, B. D. 2002 *Modern Applied Statistics with S,* New York, NY, Springer.

WAGNON, J. L., BUNTON-STASYSHYN, R. K. & MEISLER, M. H. 2016. Mutations of Sodium Channel SCN8A (Nav1.6) in Neurological Disease. *In:* PITT, G. S. (ed.) *Ion Channels in Health and Disease.* San Diego: Academic Press.

WRIGHT, M. N. & ZIEGLER, A. 2017 ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. . *Stat. Softw.,* 77**,** 1–17.

ZOU, Q. & MA, Q. 2020. The application of machine learning to disease diagnosis and treatment. *Math Biosci,* 320**,** 108305.
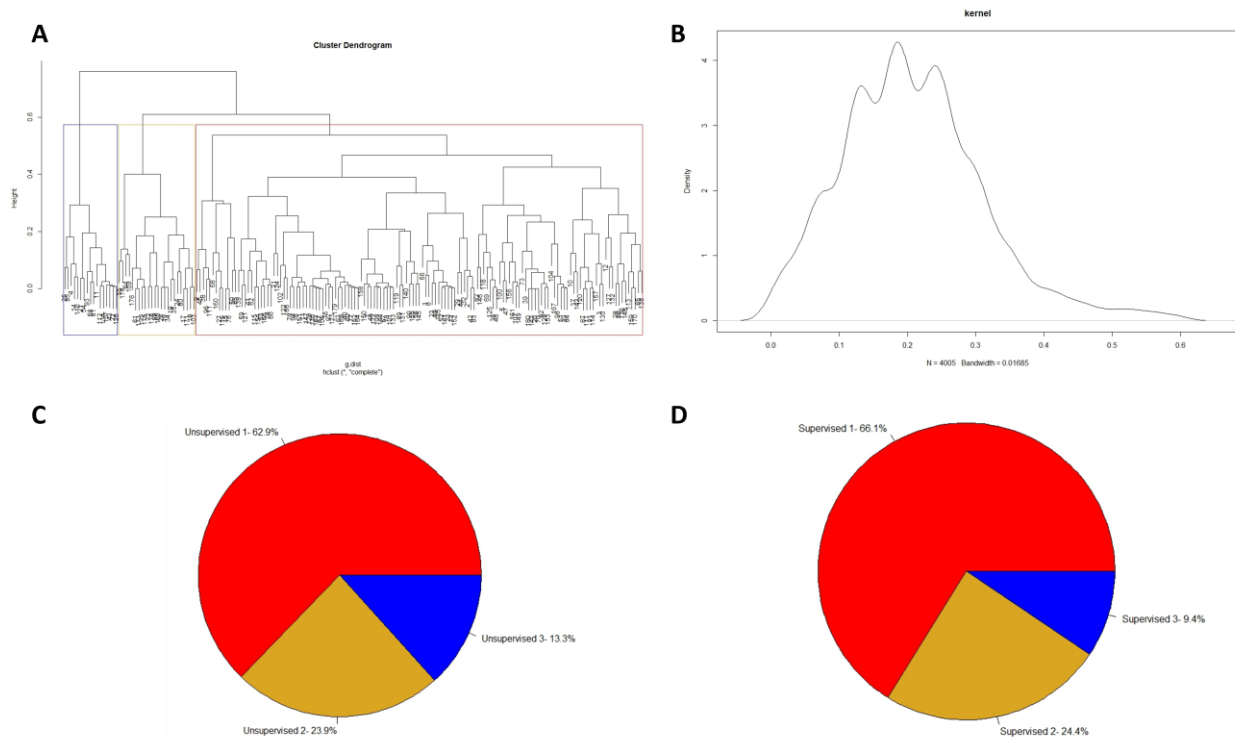
Figures and Tables

Fig. 1. The *SCN8A* GOF population can be split into three groups, with strong imbalance. (A) Hierarchical cluster analysis showing the cutoff for optimal number of groups. Hierarchical clustering used the complete agglomeration method for the optimal dendrogram. Cluster U1 (*blue,* n=24), Cluster U2 (*yellow,* n=43), and Cluster U3 (*red,* n=113) are marked in their respective colored boxes. *(B) Kernel density plot. Peaks around* 0.2 show 3 optimal clusters. (C) Distribution of patients in clusters assigned by hierarchical clustering in the *Unsupervised Approach*. (D) Distribution of patients in clusters assigned by clinical and researcher guidelines in the *Supervised Approach.*
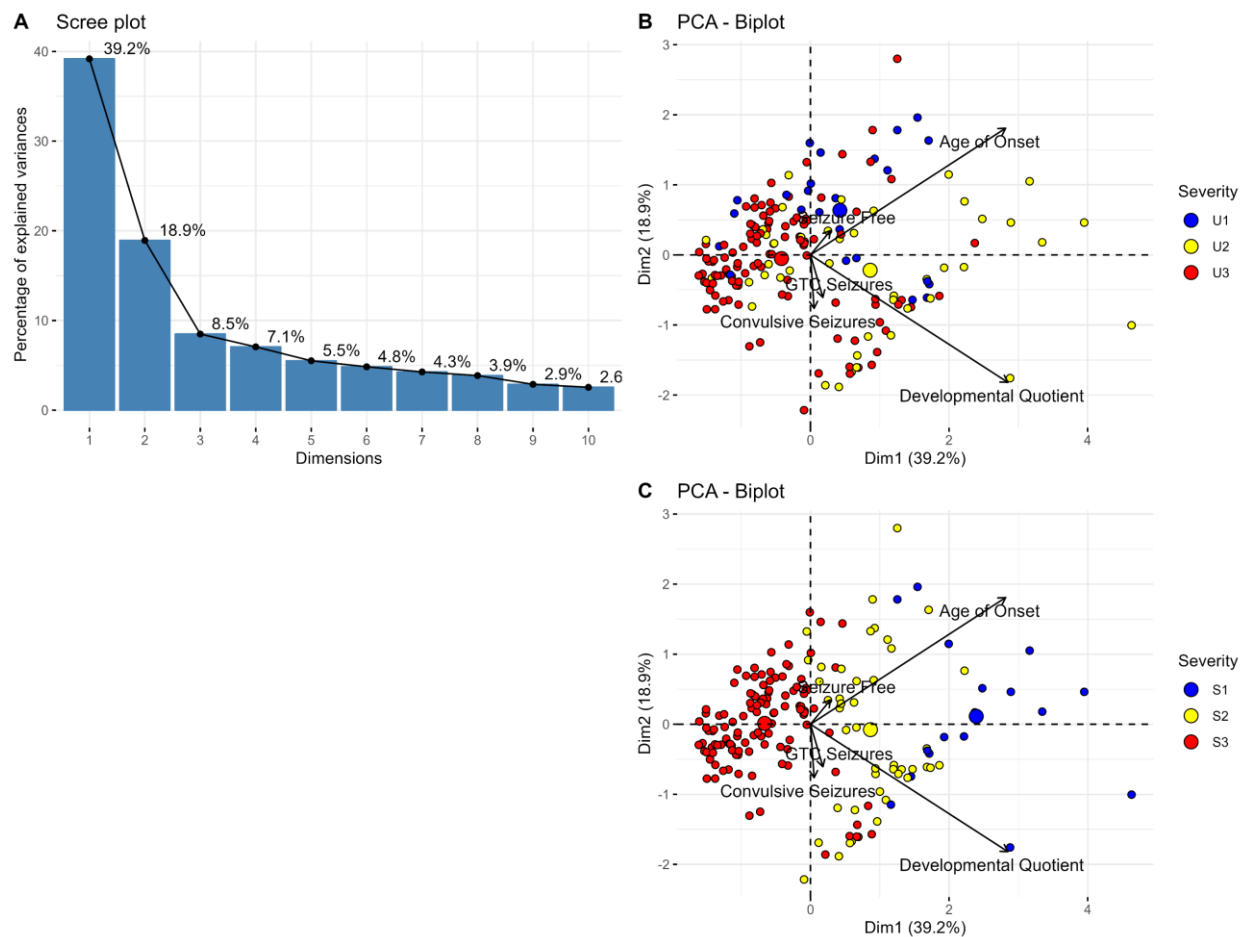
Fig. 2. Dimension reduction of *SCN8A* GOF population shows limited separation between groups.

(A) Scree plot showing percentage of variance explained in each dimension using *Unsupervised Clusters*. (B) 2-dimensional PC plot showing distribution of Cluster U3 (*red*, n=113), Cluster U2 (*yellow,* n=43), and Cluster U1 (*blue,* n=24). The five features with highest contribution to variance are shown as vectors in *black*. (C) 2-dimensional PC plot showing distribution of Cluster S3 (*red,* n=119), Cluster S2 (*yellow,* n=44), and Cluster S1 (*blue,* n=17). The five features with highest contribution to variance are shown as vectors in *black*.
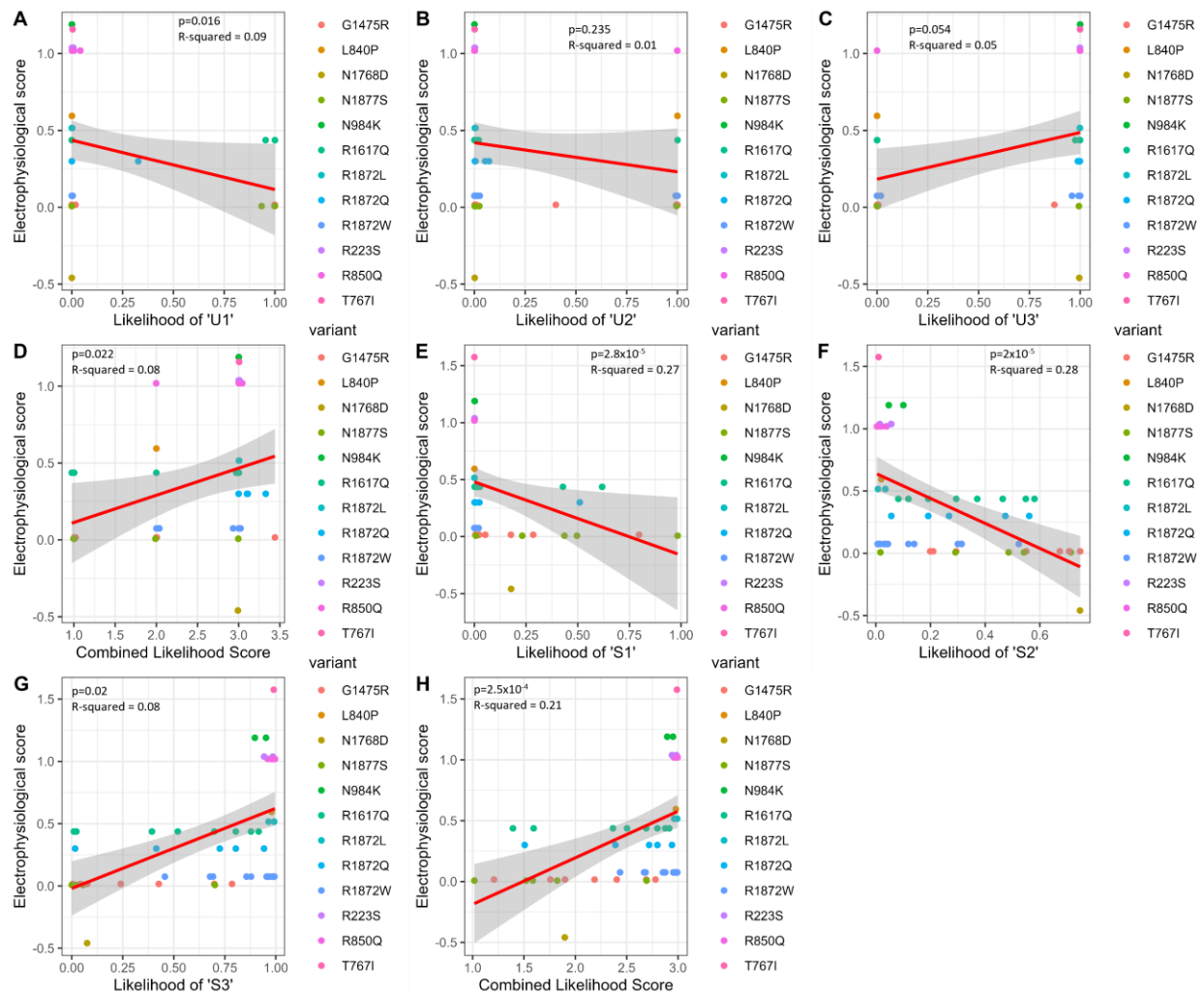
Fig. 3. Linear regression between group probability and variant electrophysiological score suggests biological relevance.

Linear regression between group probability and electrophysiological score for (A-D) *Unsupervised Stacked* and (E-H) *Supervised p-ORM*. Electrophysiological score as a function of probability of classification as (A) U1, (B) U2, (C) U3, (D) combined probability score from *Unsupervised Stacked* model, (E) S1. (F) S2, (G) S3, and (H) combined probability score from the *Supervised p-ORM* model. Variants included for *Unsupervised Stacked*: p.R223S (n=2), p.T767I (n=1), p.L840P (n=1), p.R850Q (n=7), p.N984K (n=2), p.G1475R (n=7), p.R1617Q (n=8), p.N1768D (n=1), p.R1872L (n=2) , p.R1872Q (n=4) , p.R1872W (n=11), p.N1877S (n=5). Variants included for *Supervised p-ORM:* p.R223S (n=3), p.T767I (n=1), p.L840P (n=1), p.R850Q (n=9), p.N984K (n=2), p.G1475R (n=7), p.R1617Q (n=8), p.N1768D (n=1), p.R1872L (n=2) , p.R1872Q (n=5) , p.R1872W (n=12), p.N1877S (n=5).

Table 1. Descriptive statistics of study population.

Feature	Total Population	Unsupervised			Supervised		
		U1	U2	U3	S1	S2	S3
Total Patients	180	24	43	113	17	44	119
Developmental Quotient	*36.84 (39.87)*	36.45 (33.10)	65.6 (47.26)	26.63 (32.60)	*102.07 (38.18)*	*64.04 (33.59)*	*18.08 (23.72)*
Seizure History	100%	100%	100%	100%	100%	100%	100%
Age at Onset	*3.64 (2.86)*	5.26 *(2.71)*	4.83 (3.54)	2.88 (2.28)	*8.32 (3.09)*	*5.2 (2.60)*	*2.34 (1.72)*
Seizure Freedom	16.1%	100%	11.6%	0%	35.3%	27.3%	9.2%
Developmental Delay	80.6%	100%	23.3%	98.2%	58.8%	77.3%	84.9%
Initial Seizure Type(s)							
Absence	8.9%	0%	2.3%	13.3%	0%	9.1%	10.1%
Generalized Tonic-Clonic	56.7%	29.2%	65.1%	59.3%	58.8%	56.8%	56.3%
Atonic	4.4%	0%	2.3%	6.2%	0%	4.5%	5.0%
Focal	34.4%	16.7%	18.6%	44.2%	17.6%	40.9%	34.5%
Convulsive	66.1%	45.8%	72.1%	68.1%	64.7%	68.2%	65.5%
Myoclonic	13.3%	0%	4.7%	19.5%	0%	9.1%	16.8%
Infantile Spasms	23.9%	20.8%	7.0%	31.0%	11.8%	18.2%	27.7%
Tonic	28.3%	8.3%	14.0%	38.1%	11.8%	22.7%	32.8%

Descriptive statistics of the complete study population (*Column 1*), each of the *Unsupervised Clusters* (*Columns 2-4*), and each of the *Supervised Clusters* (Columns 5-7). Developmental quotient and age at onset show mean (standard deviation) while all other features show the percentage of subgroup reporting each feature.

Table 2. Selected coefficients following penalization.

Unsupervised (λ=18)		Supervised (λ=14)	
Feature	*Coefficient*	*Feature*	*Coefficient*
Seizure Freedom	9.712	Age at Onset	0.745
Convulsive Seizures	0.838	Seizure Freedom	0.707
Developmental Quotient	0.010	Developmental Quotient	0.055
Age at Onset	0.002	Tonic Seizures	-0.005
Infantile Spasms	-0.386	Convulsive Seizures	-0.065
Myoclonic Seizures	-1.080	Bilateral Tonic-Clonic Seizures	-0.132
Absence Seizures	-1.159		
Focal Seizures	-1.254		
Tonic Seizures	-1.321		
Developmental Delay	-3.262		
Best Lambda index is included in parentheses.			

Table 3. Confusion matrix of predictions from penalized ordinal regression models.

Unsupervised				Supervised			
TRUE	**Predicted**			**TRUE**	**Predicted**		
	U1	U2	U3		S1	S2	S3
U1	24	0	0	S1	12	4	1
U2	5	27	11	S2	2	32	10
U3	0	0	113	S3	0	8	111
		Error= 0.089				Error= 0.139	

Table 4. Performance statistics of Stacked Models across three trials.

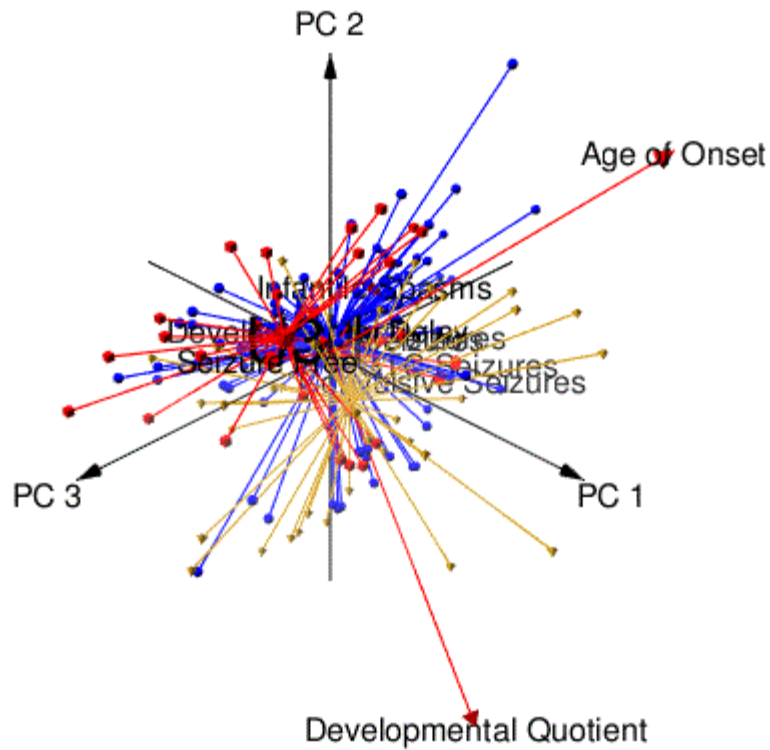| Fitness Value	Unsupervised		Supervised	
	Classification	Probability	Classification	Probability
Mean Absolute Error	0.06 (0.01)	0.04 (0.03)	0.37 (0.09)	0.35 (0.07)
Root-Mean-Square Error	0.16 (0.01)	0.15 (0.07)	0.54 (0.07)	0.52 (0.08)
Percentage of Correct Classification	94.8 (0.81)	95.2 (2.70)	81.0 (4.58)	81.0 (1.00)
Data is presented as mean (standard deviation).				

Fig. S1. 3-dimensional principal component plot of *Unsupervised Approach.* Cluster U3 (*green*), U2 (*blue*), and U1 (*yellow*) are shown with contribution from the top 5 features as vectors.

**Fig. S2. 3-dimensional principal component plot of S*upervised Approach.* Cluster S3 (*green*), S2 (*blue*), and S1 (*yellow*) are shown with contribution from the top 5 features as vectors.
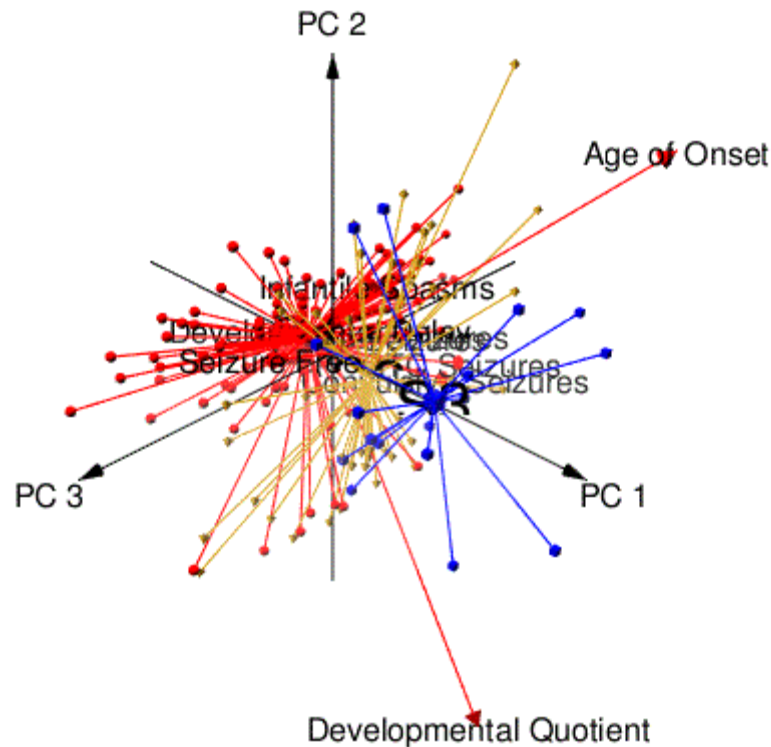
Table S1. Justification for variant inclusion in this study. Each patient's variant was verified as being pathogenic or likely pathogenic by cross-referencing patient genetic report, the ClinVar database, and *SCN8A* literature. When variant was unable to be verified using these methods, the ACMG guidelines [27] on classifying a variant as pathogenic or likely pathogenic were followed. Justification for variants being classified as GOF is included for each variant, using either known electrophysiological studies or the LOF Classifier [16]. In instances where a patient was unable to be verified using these methods, justification for being GOF is included based on clinical features and response to medications.

Available for download at
https://journals.biologists.com/bio/article-lookup/doi/10.1242/bio.060286#supplementary-data