Published in final edited form as:

Epilepsia. 2017 July ; 58(7): 1190–1198. doi:10.1111/epi.13798.

Comparison and optimization of *in silico* algorithms for predicting the pathogenicity of sodium channel variants in epilepsy

Katherine D. Holland, MD, PhD¹, Thomas M. Bouley², and Paul S. Horn, PhD¹

¹Departments of Pediatrics and Neurology, University of Cincinnati College of Medicine and Division of Child Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

²McMicken College of Arts and Sciences, University of Cincinnati, Cincinnati OH

SUMMARY

Objective—Variants in neuronal voltage gated sodium channel α -subunits genes *SCN1A*, *SCN2A*, and *SCN8A* are common in early-onset epileptic encephalopathies and other autosomal dominant childhood epilepsy syndromes. However, in clinical practice missense variants are often classified as variants of uncertain significance when missense variants are identified but heritability cannot be determined. Genetic testing reports often include results of computational tests to estimate pathogenicity and the frequency of that variant in population-based databases. The objective of this work was to enhance clinicians' understanding of results by (1) determining how effectively computational algorithms predict epileptogenicity of sodium channel (SCN) missense variants; (2) optimizing their predictive capabilities; and (3) determining if epilepsy-associated SCN variants are present in population based databases. This will help clinicians better understand results of indeterminate SCN test results in people with epilepsy.

Methods—Pathogenic, likely pathogenic, and benign variants in SCNs were identified using databases of sodium channel variants. Benign variants were also identified from population-based databases. Eight algorithms commonly used to predict pathogenicity were compared. In addition logistic regression was used to determine if a combination of algorithms could better predict pathogenicity.

Results—Based on American College of Medical Genetic Criteria, 440 variants were classified as pathogenic or likely pathogenic and 84 were classified as benign or likely benign. Twenty-eight variants previously associated with epilepsy were present in population-based gene databases. The output provided by most computational algorithms had a high sensitivity but low specificity with an accuracy of 0.52–0.77. Accuracy could be improved by adjusting the threshold for pathogenicity. Using this adjustment, the M-CAP algorithm had an accuracy of 0.90 and a combination of algorithms increased the accuracy to 0.92.

Corresponding Author: Katherine D. Holland, MD, PhD, Division of Child Neurology, ML #2015, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229 USA, katherine.holland@cchmc.org, 513-636-4222 (phone), 513-636-1888 (fax).

CONFLICTS OF INTEREST

Katherine Holland has received support from the National Institutes of Health. The remaining authors have no conflicts of interest. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

Significance—Potentially pathogenic variants are present in population-based sources. Most computational algorithms overestimate pathogenicity; however, a weighted combination of several algorithms increased classification accuracy to over 0.90.

Keywords

SCN1A; epileptic encephalopathy; *SCN2A*; *SCN8A*

INTRODUCTION

At least 10–30% of early onset epileptic encephalopathies (EOEE) result from single gene mutations^{1–4}. As a result, screening for mutations in multiple genes simultaneously or performing whole exome sequencing is common in the clinical evaluation of children with EOEE⁵. Mutations in voltage-gated sodium channel (SCN) α -subunit genes are the most commonly identified cause of EOEE; with sodium channel genes *SCN1A*^{1, 4} and *SCN2A*² being the most frequently identified causative genes. Mutations in the related SCN gene *SCN8A* have also been associated with EOEE^{1–3, 6}. Sodium channelopathies have a wide spectrum of epilepsy phenotypes; for example, mutations in *SCN1A* are classically associated with severe myoclonic epilepsy of infancy (SMEI; Dravet Syndrome) and a milder autosomal dominant phenotype, generalized epilepsy with febrile seizures plus (GEFS+)^{7, 8}. Approximately half of *SCN1A* mutations associated with SMEI are missense and half result in truncation of the protein. In contrast, the majority mutations associated with *SCN2A* and *SCN8A*, and with inherited *SCN1A*-associated phenotypes are missense mutations⁹. When SCN gene variants are identified in people with epilepsy, additional evaluation to determine pathogenicity is needed including testing of family members to establish the inheritance pattern; EOEE typically occurs *de novo*, whereas milder phenotypes are inherited from an affected parent. Unfortunately family members may not be available. In studies genetic of causes of EOEE and genetic registries of *SCN1A* mutations, complete parental testing was unavailable in about 10–40 % of cases^{1, 10, 11}. Therefore, in the absence of a classic clinical presentation or limited corroborative genetic testing, results of the initial genetic evaluation can produce equivocal results, termed variants of unknown significance (VUS). For clinical purposes, the American College of Medical Genetics and Genomics (ACMG) proposed that for classification of variants as likely pathogenic or likely benign a classification accuracy of 90% is needed.

One ‘gold standard’ for understanding the implications of variants is *in vitro* demonstration of deleterious effects; however, these electrophysiological studies are not clinically available. The lack of functional testing is not unique to epilepsy. To address this, computational methods to estimate the likelihood a mutation is pathogenic have been developed^{12–20}. To gain a better understanding of the pathogenicity of VUS, the purpose of this work is to determine how accurate current computational algorithms are at predicting pathogenicity of missense variants in SCN genes most commonly associated with epilepsy (*SCN1A*, *SCN2A*, and *SCN8A*) and to optimize the effectiveness of these algorithms.

METHODS

Potentially pathogenic nonsynonymous missense mutations in *SCN1A* were identified from two large datasets (Zuberi et al²¹ and <http://www.molgen.vib-ua.be/scn1amutations/Mutations/Mutations.cfm>²²) which included information about both phenotype and inheritance. Pathogenic and likely pathogenic variants in *SCN2A* were obtained from ClinVar (<http://www.ncbi.nlm.nih.gov/clinvar>), the publically accessible portion of HGMD (<http://www.hgmd.cf.ac.uk/ac/>) and from the literature^{3; 23; 24}. Pathogenic variants in *SCN8A* were identified from www.scn8a.net then classified by review of the associated literature. Only epilepsy-associated phenotypes were studied. Two other voltage-gated sodium channels, *SCN3A* and *SCN9A*, have also been implicated in epilepsy. However, for *SCN3A* the phenotypes are less distinct²⁵ and it is unclear if *SCN9A* variants are a cause of epilepsy or a modifier of seizure presentation²⁶. The number of potentially pathogenic variants in these genes is low, and some of the commercially available panels do not include these genes. Therefore *SCN3A* and *SCN9A* were not included in this analysis.

Genetic and phenotypic information on variants that appeared in multiple datasets were combined to better refine the phenotype. Variants were classified as pathogenic, likely pathogenic, variants of uncertain significance, benign or likely benign according to ACMG guidelines.²⁷ Likely benign variants were identified from the benign variants in datasets above, from the Exome Aggregate Consortium (ExAC) database (<http://exac.broadinstitute.org>)²⁸ and from the 1000 genomes dataset (<http://www.1000genomes.org>). The ExAC database is an aggregate of exome sequencing data from a variety of different sequencing projects. It contains sequences from 60,706 unrelated individuals without severe pediatric diseases. These population-based datasets have incomplete phenotyping so it is possible that some individuals with mild phenotypes such as GEFS+ or Benign Familial Neonatal Infantile Seizures (BFNIS) could be included. The threshold frequency for inclusion as a benign was set at a minor allele frequency (MAF) above the frequency of any pathogenic variant seen in the ExAC dataset (MAF = 0.000058). If a variant was present in a population-based dataset at a frequency above that of EOEE-associated variants and was present in a disease dataset but was classified as a VUS, it was included as a likely benign variant (based on ACMG criteria).

Eight computational algorithms were used to predict the pathogenicity of variants. These algorithms were selected because they are widely and freely-accessible, and provided both categorical and continuous results. The results of some are reported on clinically used epilepsy gene panels (e.g. GeneDx) and have been used in similar studies of cancer and cardiac diseases^{29; 30}. Computations were done using the web-based platforms shown in Table 1. PolyPhen-2 and MutationAssessor3 used graded scales. Initial results from these algorithms were dichotomized. For PolyPhen-2, probably and possibly damaging variants were classified as pathogenic and benign variants were classified as non-pathogenic. PolyPhen-2 has two different models (HumDiv and HumVar). The HumDiv model was selected because it was the algorithm used in analogous cardiac and cancer studies^{29; 30} and had fewer intermediate results than the HumVar model. For MutationAssessor3, high and medium impact variants were classified as pathogenic and low and neutral impact variants were classified as non-pathogenic. This categorization was used previously in comparative

studies of these methodologies^{29; 30} and in the development of the original algorithm¹⁹. One algorithm (M-CAP) reported scores only if the variant was an uncommon variant, so for this study an M-CAP score of zero was used for those common variants. Only variants where all programs provided an output were used in the analysis. Sensitivity, specificity, and classification accuracy were calculated for each algorithm. Classification accuracy was defined here as the average of the sensitivity and specificity. For optimization of algorithms, classification accuracy was selected over the ability of an algorithm to correctly classify individual variants (correct decision) because the correct decision approach would be biased by a greater number of pathogenic SCN variants compared to benign variants.

Several approaches to optimize the classification accuracy of algorithms were used. First the categorical output of the two algorithms that provided graded output was refined. For PolyPhen-2 and Mutation Assessor3, location of the variant within the resultant protein was considered to classify the intermediate categories. Based on the frequency of pathogenic variants in proportion to the size of functional domains in *SCN1A*, Zuberi and colleagues demonstrated that nonsynonymous missense mutations occur more frequently in some regions (N-terminus, transmembrane segments S3, S4, S5, and S6 and the pore loop connecting transmembrane segments S5 and S6) of *SCN1A* than in others (C-terminus, transmembrane segments S1, and S2, cytoplasmic linkers, and linkers between the transmembrane region other than the pore)²¹. For PolyPhen-2 the possibly damaging variants were classified as pathogenic if they occurred in regions with a greater than average pathogenic rate and non-pathogenic if in regions with a lower than average pathogenic rate. The medium impact variants in MutationAssessor3 were similarly categorized. A second method was to determine the threshold value of the continuous variables which maximized classification accuracy.

To improve classification, continuous responses from the algorithms were considered as possible explanatory variables in a logistic regression model. The backward elimination variable selection procedure was used to derive a parsimonious model that would best predict the classification of variants. The criterion used was classification accuracy. The resulting model produced a fitted value for probability of disease that we will refer to as the SCN Index. Comparisons of SCN Indices between groups were done using the Mann-Whitney-U rank sum test and the Bonferonni correction was applied to correct for multiple comparisons (15 comparisons).

A validation set was identified from different data sources. ClinVar and HGMD were used for *SCN1A* and *SCN8A*. Additional *SCN1A* variants were also identified from <http://www.gzneurosci.com/scn1adatabase/index.php>. *SCN2A* variants included in the validation were variants in the professional version of HGMD that were not in the publically accessible version of HGMD. Variants that were in the original dataset were excluded from the validation dataset. Because the phenotype and inheritance were not as readily available in all the sources used for validation, ACMG classification could not be used for all entries; therefore, the variants were classified according to the attribution in the particular datasets. For comparing the validation to the original datasets, the pathogenic and likely pathogenic variants in the original were combined because it is possible the validation dataset contained both pathogenic and likely pathogenic variants. The classification accuracy, sensitivity and

specificity were calculated using the thresholds optimized from the original dataset. In addition, variants in two epilepsy-associated voltage-gated potassium channel genes, *KCNQ2* and *KCNQ3* were identified from ClinVar (accessed 9/23/2016).

The datasets were accessed between 8/15/2015 and 9/25/2016 and websites for predictive algorithms were accessed between 8/15/2015 and 1/28/2017. SAS ® version 9.3 (SAS Institute, Inc., Cary, NC) and Origin 8.1 were used for statistical analyses. This study was exempt from IRB review.

RESULTS

From the initial disease datasets 624 unique missense variants were identified (supplementary Table 1). One was excluded because it interfered with the initiation of the protein and the full battery of *in silico* testing could not be performed and two excluded because results were unavailable from one algorithm (M-CAP). Applying ACMG criteria, 440 were categorized as pathogenic or likely pathogenic (*SCN1A* 343; *SCN2A* 62; *SCN8A* 35), 92 as VUS (*SCN1A* 45; *SCN2A* 41; *SCN8A* 6), and 84 as benign (*SCN1A* 44; *SCN2A* 24; *SCN8A* 16). VUS were most often those identified in the datasets that did not have phenotypes typically associated with sodium channelopathies and had incomplete characterization of inheritance.

Combining the ExAC database, 1000Genomes database and benign variants from Zuberi et al, 913 variants were identified (supplementary Table 2). The majority of those in ExAC occurred only once in the approximately 120,000 alleles sequenced (n=583). Included in the ExAC database were three *SCN1A* variants that were in the initial disease datasets identified as SMEI-spectrum phenotype with *de novo* inheritance and one with a GEFS+ phenotype. The most frequently seen SMEI-associated missense variant in ExAC (*SCN1A* H127D²¹) occurred in seven individuals (minor allele frequency = 0.0058%). In addition, a variant of *SCN2A* (R188W) which had demonstrated altered channel function in *in vitro* electrophysiological studies, was also seen in the ExAC dataset, occurring in 3 individuals (MAF 0.0025%). In the absence of the ExAC dataset these would have been classified as pathogenic. One *SCN1A* variant (R1596C) continued to be classified as pathogenic based on ACMG criteria despite its presence in the control dataset (due to additional supporting evidence of pathogenicity) and the others met criteria for likely pathogenic. Of note, although frame-shift variants were not included in this analysis, the ExAC database also contained a frame-shift *SCN1A* variant (R1912KfsX32; one allele) that had been previously reported *de novo* in a patient with an SMEI spectrum phenotype²¹.

Only one variant in each sodium channel occurred in more than 1% of the genotypes and the others were rare with a minor allele frequency of 0.5% or less. Excluding those in ExAC that occurred at a frequency below that associated with likely pathogenic variants, only 84 benign or likely benign variants were identified. Among these 84, based on published literature and ACMG criteria, eight would have been classified as VUS for epilepsy if they had not appeared in the ExAC database. One variant was associated with familial hemiplegic migraine.

Without optimization, the classification accuracy of predication for various algorithms ranged from 0.52 to 0.77 (Table 2). All algorithms except for FATHMM Unweighted had sensitivities of over 0.9; however most had poor specificity; and only one (FATHMM Unweighted) had a specificity above 0.7. PolyPhen-2 HumDiv and MutationAssessor3 had the highest accuracy (0.77). PolyPhen-2 HumVar had a similar classification accuracy as PolyPhen-2 HumDiv (0.78). Weighing the output of intermediate findings of MutationAssessor3 and PolyPhen-2 HumDiv for locations within mutation hotspots in *SCN1A* produced modest improvements in classification accuracy. Adjustment of the threshold value to predict pathogenicity increased the classification accuracy (to 0.77–0.9). This resulted from improvements in specificity (all > 0.7). With the adjusted threshold, the most accurate algorithm, M-CAP, had a classification accuracy of 0.9.

The most effective optimization strategy was a logistic regression model where the probability of disease was modeled. The prediction algorithms were the independent variables and a backward elimination procedure was used to remove non-statistically significant variables. The resulting combination of prediction algorithms consisted of Condel 2.0 (Odds, ratio, OR, = 1.10 for each 0.01-unit increase, 95% confidence interval, CI, = 1.02 – 1.19, p-value = 0.018), FATHMM (Weighted) (OR = 1.70 for each 1-unit decrease, 95% CI = 1.01 – 2.86, p-value = 0.046), and M-CAP (OR = 1.57 for each 0.1-unit increase, 95% CI = 1.32 – 1.87, p-value < 0.0001). The result of the equation, with the variables in their original units (see below), yielded a probability, thus, a score between 0 and 1. At a threshold of 0.782 the classification accuracy for identification of a variant as epilepsy associated versus non-epilepsy associated was 92% (sensitivity 94%; specificity 89%).

$$SCN \text{ Index} = 1 / [1 + \exp(9.1439 + 0.5312 * \text{FATHMM weighted} - 9.578 * \text{Condel} - 4.5022 * \text{M-CAP})]$$

The resulting ROC curve had an area of 0.954. Further, the Hosmer-Lemeshow goodness-of-fit test showed no lack-of-fit (p-value = 0.82). By comparison, while the area under the ROC for PolyPhen-2 HumDiv alone was 0.85, it showed significant lack-of-fit with the p-value for the Hosmer-Lemeshow test = 0.0001.

The SCN Index predicted that 411 of 440 (93.4%) variants classified according to ACMG as pathogenic or likely pathogenic would be pathogenic, only 44 of 92 (47.8%) of VUS would be categorized as pathogenic, and 9 of 84 (10.7%) categorized as benign were predicted to be pathogenic. Consistent with this, values of the SCN Index were highest in the pathogenic group and were progressively lower in less clearly pathological categories (Figure 1 and Table 3). In the cases where complete inheritance information was present, SCN Index was significantly greater in pathogenic and likely pathogenic variants that occurred *de novo* compared to those that were inherited (p = 0.001). The SCN Index was higher for *SCN1A* than *SCN2A* (p < 0.001) and there was a trend toward higher SCN Index with *SCN1A* compared to *SCN8A* but this did not reach statistical significance. This may represent differences in the classification used in the different data sources used for each channel.

DISCUSSION

Early onset epileptic encephalopathies are characterized by difficult to control seizures and developmental regression or retardation with an onset before the first year of life. As genetic causes have become increasingly recognized, sequencing of multiple genes simultaneously is frequently undertaken as part of the evaluation of imaging negative EOEE. However, our ability to do genetic analysis has progressed faster than our ability to interpret the results. The pathogenicity of *de novo* null variants (such as nonsense and frameshift variants) requires little supporting evidence²⁷ but for missense variants the criteria to establish pathogenicity rely on supportive genetic data, functional evidence, or both. Unfortunately, this supportive evidence is often not available.

Initial reports of *SCN1A* mutations associated with epilepsy appeared more than a decade ago³² and since then hundreds of different pathogenic variants in *SCN1A* have been identified. More recently mutations in *SCN2A* and *SCN8A* have also been identified in people with EOEE and families with inherited epilepsies. Consequently, it is possible to determine how computer algorithms designed to predict the effect of a missense variant on protein function in general perform more specifically in predicting the epileptogenicity of SCN variants. It is also possible to optimize the categorization of SCN variants.

Without optimization, the algorithms studied here all had a relatively low specificity and thus overestimated the potential pathogenicity of SCN variants. The sensitivity of two algorithms was improved by weighting the intermediate interpretations according to the location within the protein by assigning greater pathogenicity to those variants occurring in the N-terminus, pore region (S5, S5–S6, S6), voltage sensor region (S4), and transmembrane segment S3. This approach requires knowledge of the tertiary structure of the ion channel. While this information is available, it is often not provided in clinical reports. Another limitation is that it assumes that the proteins encoded by *SCN2A* and *SCN8A* have the same distribution of pathogenic variation as *SCN1A*.

Strategies that optimized the threshold value defining pathogenicity improved the performance of all *in silico* algorithms so that they all had positive predictive values over 0.9 but most algorithms still overestimated the potential deleterious impact of variants. However, the use of optimized paradigms is only helpful if the raw output is available. This is often not the case when reviewing commercial genetic testing reports. Nevertheless, the most accurate simple algorithm for estimation of pathogenicity of sodium channel variants was M-CAP, after the threshold for pathogenicity was adjusted to a level of 0.63. The most accurate optimization method overall, calculation of the SCN Index, is the most difficult because it includes continuous output of multiple algorithms. A website is available to make the SCN Index easily accessible (paste <https://www.cincinnatichildrens.org/service/c/epilepsy/sn1a> into an internet browser). In addition, SCN Indices and classifications for most predicted *SCN1A*, *SCN2A*, and *SCN8A* missense variants are listed in supplementary Table 3.

The classification accuracies for sodium channel-related epilepsies was over the 90% threshold set by the ACMG for clinical use for both the SCN Index and M-CAP when the

latter used the higher threshold value for pathogenicity. However, at least when the threshold optimized for sodium channels was applied to voltage-gated potassium channels, none of the paradigms reached the 90% accuracy proposed by the ACMG.

A major weakness in this approach was the difficulty defining variants as benign or pathogenic. This is explained partially by the lack of available phenotyping in population databases. Surprisingly, though variants associated with severe sporadic epilepsy are under extreme selection bias and as such should not be seen in population datasets of adults free of severe childhood disorders (such as ExAC), these did contain pathogenic variants associated with EOEE. Though relatively rare, some are seen in multiple unrelated persons and have also been identified in unaffected family members of epileptic subjects³³. Moreover, at least one variant found in the ExAC dataset had demonstrated deleterious effects on channel function *in vitro*³⁴. Though participants are free of severe pediatric diseases, which would include EOEE; people with milder phenotypes such as GEFS+ could be represented. Alternatively, there may be other factors that influence the expression of variants in mildly affected individuals. One approach that could address this issue would be to restrict analysis only to those variants that have functional studies. However, only a limited number of variants have been studied; therefore, this approach is currently not feasible. Further optimization of prediction programs would require greater availability of both pathogenic and benign variants. While the number of pathogenic variants in sodium channels continues to increase; identification of additional truly benign variants is slower but may become available if large population-based databases with clinical phenotyping become available. The *SCN1A*, *SCN2A*, and *SCN8A* genes are relatively intolerant to mutations, increasing the likelihood that variants in these genes lead to an epilepsy phenotype⁴. This intolerance also means that relatively few benign variants are available. Most of the prediction algorithms used here are based on conservation of genes compared to related genes. As a result of relative intolerance to change and high conservation of the voltage-gated sodium channel genes, as new gene sequences become available and are incorporated into the prediction programs, the prediction scores for these genes are unlikely to change dramatically.

From the clinical prospective, identification of a missense *SCN1A*, *SCN2A*, or *SCN8A* variant in an infant with severe epilepsy should raise suspicion, even if that variant is present at a very low frequency in large population databases. In cases when further confirmatory information such as when parental DNA is unavailable or the clinical phenotype is atypical, certain *in silico* prediction algorithms such as M-CAP or the SCN Index can be helpful. Because the SCN Index tends to be lower in milder phenotypes, a high SCN Index in a mild epilepsy phenotype is also suspicious.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grant R01-NS062756. The authors would also like to thank Yonggen Song for assistance in locating the *SCN8A* variants in *scn8ainfo.net*, Ravindra Arya for assistance in generating Figure 1, and Anna Byars for reviewing the manuscript.

References

1. Carvill GL, Heavin SB, Yendle SC, et al. Targeted resequencing in epileptic encephalopathies identifies de novo mutations in *CHD2* and *SYNGAP1*. *Nat Genet.* 2013; 45:825–830. [PubMed: 23708187]
2. Trump N, McTague A, Brittain H, et al. Improving diagnosis and broadening the phenotypes in early-onset seizure and severe developmental delay disorders through gene panel analysis. *J Med Genet.* 2016; 53:310–317. [PubMed: 26993267]
3. Mercimek-Mahmutoglu S, Patel J, Cordeiro D, et al. Diagnostic yield of genetic testing in epileptic encephalopathy in childhood. *Epilepsia.* 2015; 56:707–716. [PubMed: 25818041]
4. Epi KC, Allen AS, et al. Epilepsy Phenome/Genome P. De novo mutations in epileptic encephalopathies. *Nature.* 2013; 501:217–221. [PubMed: 23934111]
5. Myers CT, Mefford HC. Advancing epilepsy genetics in the genomic era. *Genome Med.* 2015; 7:91. [PubMed: 26302787]
6. Veeramah KR, O'Brien JE, Meisler MH, et al. 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.* 2012
7. Goldberg-Stern H, Aharoni S, Afawi Z, et al. Broad phenotypic heterogeneity due to a novel *SCN1A* mutation in a family with genetic epilepsy with febrile seizures plus. *J Child Neurol.* 2014; 29:221–226. [PubMed: 24257433]
8. Marini C, Scheffer IE, Nabbout R, et al. The genetics of Dravet syndrome. *Epilepsia.* 2011; 52(Suppl 2):24–29. [PubMed: 21463275]
9. Meisler MH, Helman G, Hammer MF, et al. *SCN8A* encephalopathy: Research progress and prospects. *Epilepsia.* 2016; 57:1027–1035. [PubMed: 27270488]
10. Depienne C, Trouillard O, Saint-Martin C, et al. Spectrum of *SCN1A* gene mutations associated with Dravet syndrome: analysis of 333 patients. *J Med Genet.* 2009; 46:183–191. [PubMed: 18930999]
11. Meng H, Xu HQ, Yu L, et al. The *SCN1A* mutation database: updating information and analysis of the relationships among genotype, functional alteration, and phenotype. *Hum Mutat.* 2015; 36:573–580. [PubMed: 25754450]
12. Shihab HA, Gough J, Mort M, et al. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics.* 2014; 8:11. [PubMed: 24980617]
13. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat.* 2013; 34:57–65. [PubMed: 23033316]
14. Schwarz JM, Cooper DN, Schuelke M, et al. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods.* 2014; 11:361–362. [PubMed: 24681721]
15. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011; 39:e118. [PubMed: 21727090]
16. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 2009; 4:1073–1081. [PubMed: 19561590]
17. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet.* 2011; 88:440–449. [PubMed: 21457909]
18. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010; 7:248–249. [PubMed: 20354512]
19. Choi Y, Sims GE, Murphy S, et al. Predicting the functional effect of amino acid substitutions and indels. *PLoS One.* 2012; 7:e46688. [PubMed: 23056405]

20. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet.* 2016
21. Zuberi SM, Brunklaus A, Birch R, et al. Genotype-phenotype associations in SCN1A-related epilepsies. *Neurology.* 2011; 76:594–600. [PubMed: 21248271]
22. Lossin C. A catalog of SCN1A variants. *Brain Dev.* 2009; 31:114–130. [PubMed: 18804930]
23. Howell KB, McMahon JM, Carvill GL, et al. SCN2A encephalopathy: A major cause of epilepsy of infancy with migrating focal seizures. *Neurology.* 2015; 85:958–966. [PubMed: 26291284]
24. Herlenius E, Heron SE, Grinton BE, et al. SCN2A mutations and benign familial neonatal-infantile seizures: the phenotypic spectrum. *Epilepsia.* 2007; 48:1138–1142. [PubMed: 17386050]
25. Vanoye CG, Gurnett CA, Holland KD, et al. Novel SCN3A Variants Associated with Focal Epilepsy in Children. *Neurobiol Dis.* 2013
26. Mulley JC, Hodgson B, McMahon JM, et al. Role of the sodium channel SCN9A in genetic epilepsy with febrile seizures plus and Dravet syndrome. *Epilepsia.* 2013; 54:e122–126. [PubMed: 23895530]
27. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015; 17:405–424. [PubMed: 25741868]
28. ExomeAggregationConsortium. Analysis of protein-coding genetic variation in 60,706 humans. 2015 bioRxiv. ePub.
29. Leong IU, Stuckey A, Lai D, et al. Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. *BMC Med Genet.* 2015; 16:34. [PubMed: 25967940]
30. Martelotto LG, Ng CK, De Filippo MR, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biol.* 2014; 15:484. [PubMed: 25348012]
31. Misra SN, Kahlig KM, George AL Jr. Impaired NaV1.2 function and reduced cell surface expression in benign familial neonatal-infantile seizures. *Epilepsia.* 2008; 49:1535–1545. [PubMed: 18479388]
32. Claes L, Del-Favero J, Ceulemans B, et al. De novo mutations in the sodium-channel gene SCN1A cause severe myoclonic epilepsy of infancy. *Am J Hum Genet.* 2001; 68:1327–1332. [PubMed: 11359211]
33. Nabbout R, Gennaro E, Dalla Bernardina B, et al. Spectrum of SCN1A mutations in severe myoclonic epilepsy of infancy. *Neurology.* 2003; 60:1961–1967. [PubMed: 12821740]
34. Sugawara T, Tsurubuchi Y, Agarwala KL, et al. A missense mutation of the Na⁺ channel alpha II subunit gene Na(v)1.2 in a patient with febrile and afebrile seizures causes channel dysfunction. *Proc Natl Acad Sci U S A.* 2001; 98:6384–6389. [PubMed: 11371648]

KEY POINTS

- The output provided by commonly used computational algorithms overestimates the pathogenicity of sodium channel variants.
- Pathogenic variants of sodium channels can be found in population-based databases of gene variants.
- With adjustment of the threshold for pathogenicity, the accuracy of the M-CAP algorithm is 0.90.
- A weighted combination of algorithms can classify the pathogenicity of sodium channel variants with an accuracy of 0.92.
- The use of this algorithm may be useful for interpretation of some missense variants in the event parental DNA samples are unavailable for segregation analysis.

Figure 1. Box and whisker plots showing the distributions of SCN Indices between different classifications

All groups are different from the pathogenic group ($p < 0.001$). In an additional comparison between the inherited and *de novo* variants was also statistically significant ($p = 0.001$). The red line across the graph represents the optimized threshold above which the SCN Index predicts a variant is pathogenic. The box represents the 25–75th percentiles, bar within each box is the median; whiskers mark the outliers within 1.5 times the interquartile range; and the colored circles represent the individual data points. The exact values are listed in Table 3.

TABLE 1
Some commonly used *in silico* predictive algorithms used to predict effects of missense variants on protein function.

Prediction Algorithm	Basis	OUTPUT		Website
		Categorical (number of categories)	Continuous (Threshold value for Pathogenic classification)	
CONDEL 2.0 ¹⁷	Conservation (based on FATHMM and Mutation Assessor)	Yes (2)	Yes (0.52)	http://bg.upf.edu/famnsdb/
FATHMM (unweighted) ¹³	Species dependent conservation	Yes (2)	Yes (-3.0)	http://fathmm.biocompute.org.uk
FATHMM (weighted) ¹²	Similar to FATHMM but includes weighted consideration of disease categories	Yes (2)	Yes (-1.5)	http://fathmm.biocompute.org.uk
M-CAP ²⁰	Conservation and protein properties. Combines multiple predictive algorithms and adds additional protein and genomic level measures	Yes (2)	Yes (0.025)	http://bejerano.stanford.edu/MCAP/
Mutation Assessor rel 3 ¹⁵	Conservation within and across species	Yes (4)	Yes (1.935)	http://mutationassessor.org/
PolyPhen-2 HumDiv ¹⁸	Protein structure and conservation	Yes (3)	Yes (0.432)	http://genetics.bwh.harvard.edu/pph2/bgi.shtml
PROVEAN ¹⁹	Differences in amino acid alignment	Yes (2)	Yes (-2.5)	http://provean.jcvi.org/index.php
SIFT ¹⁶	Conservation of amino acids from closely related sequences	Yes (2)	Yes (0.05)	http://sift.jcvi.org/

TABLE 2
Accuracy and optimization of algorithms in predicting epileptogenicity of sodium channel α -subunit gene missense variants.

Prediction Algorithm	Threshold	TP	TN	FP	FN	PPV	NPV	Sensitivity	Specificity	Correct Decision	Classification Accuracy*
CONDEL 2.0	0.52	437	20	64	3	0.87	0.87	0.99	0.24	0.87	0.62
FATHMM Unweighted	-3.0	315	60	24	126	0.93	0.32	0.71	0.71	0.71	0.71
FATHMM Weighted	-1.5	440	3	81	0	0.84	1	1	0.04	0.85	0.52
M-CAP	0.025	439	6	78	1	0.85	0.86	0.99	0.07	0.85	0.54
MutationAssessor3	1.935	408	52	32	33	0.93	0.61	0.92	0.62	0.88	0.77
PolyPhen-2 HumDiv	0.432	419	50	34	22	0.92	0.69	0.95	0.60	0.89	0.77
PROVEAN	-2.5	430	32	52	11	0.89	0.74	0.97	0.38	0.88	0.68
SIFT	0.05	435	31	53	5	0.89	0.86	0.99	0.37	0.89	0.68
Optimization Method						Dichotomization of Intermediate results based on location					
MutationAssessor3 Adj	1.935	347	77	7	94	0.98	0.45	0.79	0.92	0.81	0.85
PolyPhen-2 Adj	0.432	405	59	25	36	0.94	0.62	0.92	0.70	0.88	0.81
Optimization Method						Optimization of Threshold for Classification Accuracy					
CONDEL 2.0	0.61	371	70	14	70	0.96	0.50	0.84	0.83	0.84	0.84
FATHMM Unweighted	-2.26	340	65	19	101	0.95	0.39	0.77	0.77	0.77	0.77
FATHMM Weighted	-4.15	396	64	20	45	0.95	0.59	0.90	0.76	0.88	0.83
M-CAP	0.63	404	75	9	37	0.98	0.37	0.92	0.89	0.91	0.90
MutationAssessor3	2.65	358	70	14	83	0.96	0.46	0.81	0.83	0.82	0.82
PolyPhen-2	0.98	376	65	19	65	0.95	0.50	0.85	0.77	0.84	0.81
PROVEAN	-3.62	373	67	17	68	0.96	0.50	0.85	0.80	0.84	0.82
SIFT	0.001	406	68	16	35	0.96	0.66	0.92	0.81	0.90	0.87
Optimization Method						Logistic Regression Model					
SCN Index	0.782	413	75	9	28	0.98	0.73	0.94	0.89	0.93	0.92

* Classification accuracy was defined as the average of specificity and sensitivity.
TP = true positive, TN = true negative, FP = false positive, FN = false negative; PPV= positive predictive value; NPV=negative predictive value

TABLE 3
Comparison of SCN Indices between various sodium and potassium channel classifications

	n	Median	25 th percentile	75 th percentile	Min	Max
Pathogenic/Likely Pathogenic						
All SCN	440	0.985	0.965	0.992	0.046	0.999
de novo SCN	271	0.987	0.967	0.993	0.067	0.998
Inherited SCN	75	0.974	0.881	0.987	0.101	0.999
SCN1A	343	0.974	0.987	0.993	0.046	0.999
SCN2A	62	0.955	0.874	0.982	0.101	0.998
SCN8A	35	0.980	0.931	0.987	0.704	0.996
KCNQ2 and KCNQ3	68	0.971	0.938	0.987	0.004	0.997
VUS						
All SCN	92	0.820	0.291	0.966	0.021	0.992
Benign						
All SCN	84	0.229	0.111	0.467	0.010	0.989
KCNQ2 and KCNQ3	13	0.510	0.152	0.941	0.004	0.961
Exome Aggregate Consortium all SCN						
All	911	0.299	0.123	0.751	0.002	0.997
Epilepsy	28	0.739	0.284	0.964	0.010	0.992
No Epilepsy	883	0.292	0.119	0.738	0.002	0.997
Validation SCN						
Pathogenic/Likely Pathogenic	137	0.983	0.967	0.992	0.014	0.999
Benign	6	0.294	0.208	0.613	0.015	0.630

Efficacy of optimized predictive algorithms between (1) the original dataset and an independent validation dataset, (2) pathogenic SCN missense variants from all sources and all non-epilepsy associated SCN missense variants within ExAC, and (3) potassium channels *KCNQ2* and *KCNQ3*

Table 4

Prediction Algorithm	Validation Using Separate Dataset (n=143)			SCN Pathogenic and SCN ExAC Variants (n=1458)			Application to <i>KCNQ2</i> and <i>KCNQ3</i> (N=81)		
	Sensitivity	Specificity	Classification Accuracy	Sensitivity	Specificity	Classification Accuracy	Sensitivity	Specificity	Classification Accuracy
CONDEL 2.0	0.84	0.50	0.67	0.84	0.71	0.78	1	0.36	0.68
FATHMM Unwght	0.71	0.83	0.77	0.70	0.68	0.69	0.68	0.69	0.68
FATHMM Wght	0.81	0.83	0.82	0.88	0.71	0.79	0.94	0.23	0.59
M-CAP	0.92	1	0.96	0.92	0.81	0.87	0.85	0.69	0.77
MutationAssessor3	0.83	0.50	0.67	0.81	0.73	0.77	0.69	1	0.85
PROVEAN	0.90	0.50	0.70	0.85	0.69	0.77	0.63	1	0.82
PolyPhen-2 HumDiv	0.83	1	0.92	0.84	0.68	0.76	0.82	0.77	0.80
SIFT	0.92	0.83	0.88	0.92	0.66	0.79	0.75	0.92	0.84
MutationAssessor Adj	0.85	0.83	0.84	0.80	0.85	0.83	NA	NA	NA
PolyPhen-2 Adj	0.90	0.67	0.76	0.92	0.56	0.75	NA	NA	NA
SCN Index	0.92	1	0.96	0.93	0.77	0.85	0.69	0.90	0.79