Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets

Perry Evans,[1] Chao Wu,[2] Amanda Lindy,[3] Dianalee A. McKnight,[3] Matthew Lebo,[4,5] Mahdi Sarmady,[2,6] and Ahmad N. Abou Tayoun[2,6,7]

[1]*Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA;* [2]*Division of Genomic Diagnostics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA;* [3]*GeneDx, Gaithersburg, Maryland 20877, USA;* [4]*Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine, Cambridge, Massachusetts 02139, USA;* [5]*Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA;* [6]*Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA;* [7]*Al Jalila Children's Specialty Hospital, Dubai, United Arab Emirates*

Recent advances in DNA sequencing have expanded our understanding of the molecular basis of genetic disorders and increased the utilization of clinical genomic tests. Given the paucity of evidence to accurately classify each variant and the difficulty of experimentally evaluating its clinical significance, a large number of variants generated by clinical tests are reported as variants of unknown clinical significance. Population-scale variant databases can improve clinical interpretation. Specifically, pathogenicity prediction for novel missense variants can use features describing regional variant constraint. Constrained genomic regions are those that have an unusually low variant count in the general population. Computational methods have been introduced to capture these regions and incorporate them into pathogenicity classifiers, but these methods have yet to be compared on an independent clinical variant data set. Here, we introduce one variant data set derived from clinical sequencing panels and use it to compare the ability of different genomic constraint metrics to determine missense variant pathogenicity. This data set is compiled from 17,071 patients surveyed with clinical genomic sequencing for cardiomyopathy, epilepsy, or RASopathies. We further use this data set to demonstrate the necessity of disease-specific classifiers and to train PathoPredictor, a disease-specific ensemble classifier of pathogenicity based on regional constraint and variant-level features. PathoPredictor achieves an average precision >90% for variants from all 99 tested disease genes while approaching 100% accuracy for some genes. The accumulation of larger clinical variant training data sets can significantly enhance their performance in a disease- and gene-specific manner.

[Supplemental material is available for this article.]

Comprehensive sequencing has become the cornerstone of genomic medicine and research. However, unlike previous targeted or single gene testing, multigene sequencing can yield thousands of rare variants often requiring manual clinical correlation and interpretation. Unlike synonymous (or silent) and loss-of-function (mainly nonsense, frameshift, and canonical splice site) variants for which the impact on the protein can be relatively easily predicted, novel missense variants are the most challenging to interpret, often leading to inconclusive genomic reports and leaving clinicians and families with uncertainties. On the other hand, researchers are currently incapable of studying the impact of every possible missense variant in the ∼20,000 genes of the human genome. Therefore, novel clinical-grade approaches are needed to assist clinicians and researchers in determining the pathogenicity of missense variants.

Machine learning has yielded several pathogenicity prediction tools built with variant features and previously assigned pathogenic and benign labels. Collections of labeled variant labels for classifier training and testing include the Human Gene Mutation Database (HGMD) (Stenson et al. 2009), the Leiden Open Variation Database (Fokkema et al. 2011), and ClinVar (Landrum et al. 2016). In addition, frequently occurring variants from databases like the Genome Aggregation Database (gnomAD) (Lek et al. 2016) are used as a substitute for benign variants. Variant features can describe single positions (e.g., genomic sequence context and amino acid conservation) or regions that contain the variant (e.g., protein domains and variation constraint).

Two uses of simple region features are seen in the Functional Analysis through Hidden Markov Models (FATHMM) (Shihab et al. 2013) and the Variant Effect Scoring Tool (VEST) (Carter et al. 2013). FATHMM and VEST were found to be the most important features for determining pathogenicity in an ensemble of 18 prediction scores called REVEL (Ioannidis et al. 2016). VEST distinguished disease missense variants in HGMD from high frequency (allele frequency >1%) missense variants from the Exome Sequencing Project (ESP) (NHLBI GO Exome Sequencing Project 2013, http://evs.gs.washington.edu/EVS/) using a random forest with 86 features from the SNVBox database (Wong et al. 2011). These features describe amino acid substitutions, regional amino acid composition, conservation scores, local protein structure, and annotations of functional protein sites. FATHMM scored

Corresponding author: Ahmad.Tayoun@ajch.ae

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.240994.118.

variants by their conservation in homologous sequences, weighted by the tolerance of each variant's protein family (Pfam) domain or SUPERFAMILY (Gough et al. 2001) to mutations observed in HGMD and the set of functionally neutral UniProt variants (The UniProt Consortium 2017). VEST's inclusion of functional protein sites and FATHMM's Pfam domain tolerance consideration enabled them to capture regional protein features such as domain structure and conservation, but did not capture regional tolerance to genetic variation.

Large variant collections like the Exome Aggregation Consortium (ExAC) data set (Lek et al. 2016) and gnomAD have enabled metrics that summarize purifying selection within genomic regions. Regions with high purifying selection are constrained and have less population variation than expected. Regions with low purifying selection are unconstrained and have equal or more population variation compared to expectation. With this knowledge, classifiers might flag a variant as pathogenic if it lies within a genomic region that selects against variants (Amr et al. 2017). Here, we examine three constraint metrics derived from ExAC or gnomAD. One such constraint metric is the constrained coding region (CCR) percentile, which compared observed variant counts from gnomAD to those predicted by CpG density (Havrilla et al. 2019). A similar feature called missense depletion was constructed for the missense badness, PolyPhen-2, and constraint (MPC) pathogenicity classifier of de novo missense variants (Samocha et al. 2017). MPC's missense depletion feature was measured as the fraction of expected ExAC variation that was observed in exons. Only ExAC variants with minor allele frequencies <0.1% were considered. The expected rate of rare missense variants was based on a model that used both gene and sequence context specific mutation rates (Samocha et al. 2014). An additional pathogenicity feature introduced by MPC was missense badness, which accounted for an amino acid substitution's increase in deleteriousness when it occurs in a missense-constrained region. The third constraint metric is the missense tolerance ratio (MTR), which was calculated in 31 codon windows using missense and synonymous variant frequencies from ExAC and gnomAD (Traynelis et al. 2017). MTR is the ratio of the observed missense variant fraction to the missense variant fraction calculated from all possible variants in the window when all nucleotide changes are equally likely. A variant in a low MTR region is expected to have a high chance of being pathogenic.

In this paper, we evaluate the effectiveness of region-based pathogenicity predictors in a clinical setting. We introduce three patient variant training data sets gathered from clinical sequencing panels for cardiomyopathy, epilepsy, and RASopathies. These data sets cover 17,071 patients. All variants have been manually classified by two main clinical laboratories, whose members significantly contributed to the development of the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) sequence variant interpretation guidelines (Richards et al. 2015). We use each data set to compare CCR, FATHMM, missense badness, missense depletion, MTR, and VEST, and to train disease-specific predictors. These clinical variant sets have the advantage of being consistently reviewed in a clinically sound manner and originate from focused disease studies. This allows us to explore the hypothesis that disease-specific classifiers, first introduced for smaller gene sets (Tavtigian et al. 2008; Homburger et al. 2016), are better than general genome-wide classifiers. We also introduce PathoPredictor, a disease-specific pathogenicity score trained with clinical sequencing panel variants to combine the pathogenicity scores compared here.

Results

Variants and genes studied

We focused on patient variants from three disease panels: cardiomyopathy, epilepsy, and RASopathies (Fig. 1). We also investigated the subset of epilepsy dominant genes: *CDKL5*, *KCNQ2*, *KCNQ3*, *PCDH19*, *SCN1A*, *SCN1B*, *SCN2A*, *SCN8A*, *SLC2A1*, *SPTAN1*, *STXBP1*, and *TSC1*. These genes account for a large number of epilepsy pathogenic variants and, because they follow a dominant inheritance pattern, might have distinct characteristics impacting variant prediction relative to all other epilepsy genes (see below). For each disease variant set, we compared the performance of CCR, FATHMM, missense badness, missense depletion, MTR, and VEST using panel and ClinVar variants with pathogenicity labels. We also built PathoPredictor, an ensemble classifier of pathogenicity, and tested it with variants from ClinVar not found in our disease panel variant sets. To ensure the reliability of ClinVar variant pathogenicity labels, we examined only unambiguously pathogenic or benign variants, and split ClinVar into two variant groups: all ClinVar variants and those that have been reviewed. Few of these variant collections have an equal amount of pathogenic and benign variants, with a drastic imbalance for cardiomyopathy panel variants.

Disease-specific classifier evaluation

We used disease panel and ClinVar data sets to compare pathogenicity classifiers and to train and test PathoPredictor (Fig. 2). For each disease, we used panel and ClinVar variants to build precision-recall curves using pathogenicity scores from CCR, FATHMM, missense badness, missense depletion, MTR, and VEST. These curves were summarized using average precision. To evaluate PathoPredictor, we examined each disease panel gene, trained a model using disease panel variants not found in the selected gene, and tested the model using panel and ClinVar variants from the gene. By training PathoPredictor with disease-specific variants, we collected variants that belong to genes that are more likely to share a common biological pathway and might have similar tolerance to variants.

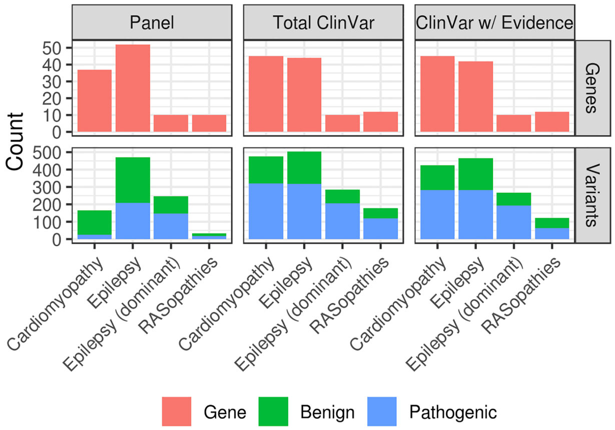

Figure 1. Study data sets. Missense variant and gene counts are shown by disease panel and ClinVar variant set. We only used ClinVar variants from panel genes and considered either any ClinVar variant (Total ClinVar), or ClinVar variants that have been reviewed (ClinVar w/Evidence). ClinVar variants were restricted to those with no conflicting pathogenicity assignments, and any genomic position from the panel data was removed from the ClinVar variant sets.

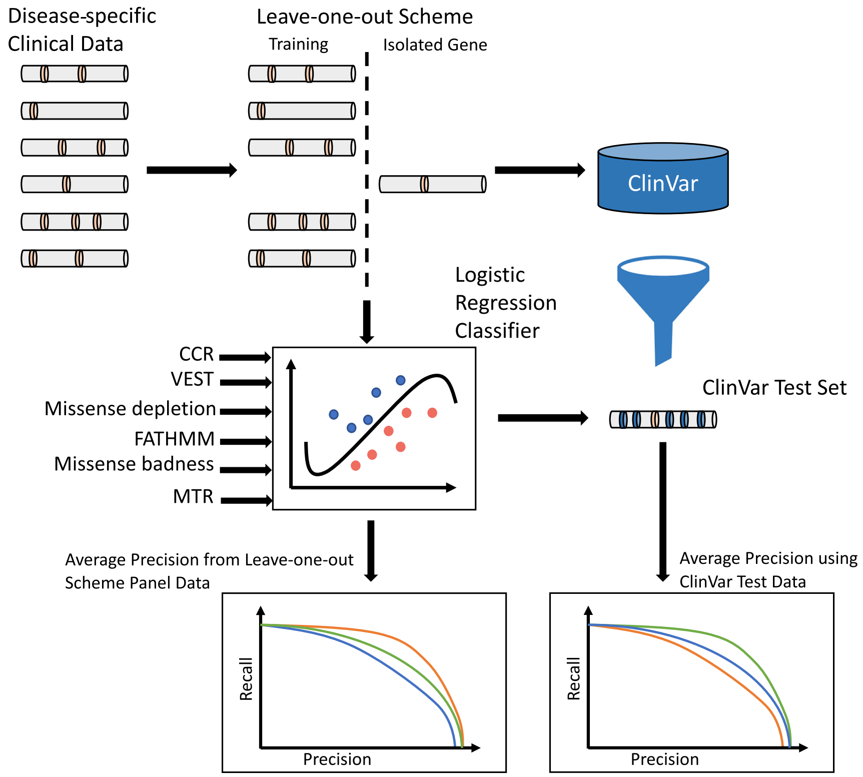
Disease-specific Clinical Data
Leave-one-out Scheme
Training
Isolated Gene
ClinVar
Logistic Regression Classifier
CCR
VEST
Missense depletion
FATHMM
Missense badness
MTR
ClinVar Test Set
Average Precision from Leave-one-out Scheme Panel Data
Average Precision using ClinVar Test Data
Recall
Precision
Recall
Precision

Figure 2. Method description. Our goal was to build disease-specific classifiers of missense variant pathogenicity using variants from clinical panels. For all genes in a disease panel, we trained a model using variants from all other genes except the gene in question and tested the model using variants from that gene of interest. We then used ClinVar variants from the gene of interest as an independent test set. Test results were summarized as average precision scores.

PathoPredictor predicts the pathogenicity of disease panel variants with an average precision higher than that obtained with any single feature (Fig. 3). The average precision of PathoPredictor is >90% for all disease panels. CCR has the highest average precision among the single features. The majority of the time, PathoPredictor's performance was significantly better than that of any single feature in the 24 comparisons made across six features and four panel variant sets (18 of the total 24 comparisons in all four panels, $P<0.05$). PathoPredictor was not significantly better than CCR for the dominant epilepsy genes, where it was expected that regional constraint would be most critical (see above). The remaining five exceptions were in the RASopathies and cardiomyopathy panels, which both had the lowest variant count (Fig. 1). However, when evaluating with more variants (below), a significant advantage of PathoPredictor over all single features was observed ($P<0.03$) (Fig. 4).

PathoPredictor's performance was next evaluated using a larger independent variant set from ClinVar (Fig. 4). For each disease panel, we found that PathoPredictor performed significantly better than any single feature when examining all ClinVar variants ($P<0.03$). PathoPredictor performed similarly when using all of ClinVar and the reviewed subset of ClinVar, achieving an average precision >95% for all variant sets. The poor average precision obtained when using VEST, FATHMM, and missense badness to predict cardiomyopathy panel variants was not replicated using ClinVar variants in cardiomyopathy panel genes. This discrepancy can be attributed to the lower number of cardiomyopathy panel variants, especially pathogenic variants (Fig. 1). PathoPredictor showed consistent results for RASopathy ClinVar and panel variants; however, given the larger number of ClinVar variants, the improved performance of PathoPredictor was now statistically significant in ClinVar ($P<0.004$).

As a further test of PathoPredictor, we trained PathoPredictor with ClinVar variants and evaluated each classifier with disease panel variants (Fig. 5). For training, we used either total ClinVar variants or those that had been reviewed, and we restricted ClinVar training variants to genes from the disease panel used for evaluation. PathoPredictor achieved an average precision of at least 90% for all evaluations. PathoPredictor performed better than each of its six features ($P<0.05$), except for missense depletion for cardiomyopathy panel variants, CCR, missense badness, and VEST for RASopathy panel variants (most likely owing to limited cardiomyopathy and RASopathy panel variants), and CCR for dominant epilepsy panel variants. These findings are consistent with the disease panel hold-one-gene-out approach in Figure 3.

Assessing the gene-wise performance of PathoPredictor is challenging because most genes have a small variant sample size. However, some genes with high variant count were found to best demonstrate the utility of PathoPredictor (Fig. 6A). When using the hold-one-gene-out approach for training and evaluation on disease panel data, PathoPredictor had an accuracy of 95% for the 27 pathogenic and 10 benign variants in *KCNQ2*. When training on panel variants and validating with ClinVar variants, PathoPredictor had a 96% accuracy for 41 pathogenic and six benign ClinVar variants in *KCNQ2*. High accuracies were also observed for *RAF1*, *SCN2A*, *SCN5A*, and *STXBP1*.

Comparing PathoPredictor with MPC and REVEL

REVEL is a state-of-art ensemble classifier of pathogenicity. Built using 18 prediction scores, it has more features than PathoPredictor, but does not contain recent genomic constraint features like CCR and missense depletion. MPC is a recent classifier of pathogenicity that was trained to combine genomic constraint features with PolyPhen-2 scores using missense pathogenic ClinVar variants and a benign variant set constructed using missense variants with 1% or higher ExAC frequencies. We used the set of de novo variants to compare PathoPredictor, MPC, and REVEL (Fig. 6B; Methods). We focused on our PathoPredictor epilepsy classifiers because they were expected to be most relevant to the neurodevelopmental and autism disorder variants from the validation de novo data set. We found that the dominant epilepsy trained PathoPredictor achieved >94% average precision, which was significantly higher than that of REVEL or MPC ($P<0.05$) (Fig. 6C). Both PathoPredictor classifiers achieved a greater average precision than REVEL ($P<0.05$) (Fig. 6C).

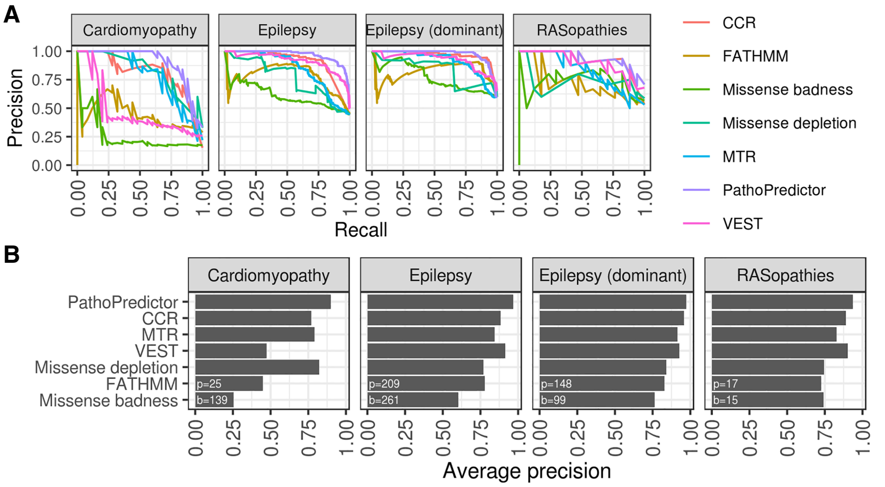

Figure 3. Disease-specific classifier performance using disease panel cross-validation. For each disease panel, we used a hold-one-gene-out approach to evaluate a logistic regression model's ability to predict pathogenicity. For all genes in a disease panel, we trained PathoPredictor using variants from all other genes and tested the model using variants from the gene of interest. Using the held-out gene variant prediction scores, we computed a precision-recall curve (*A*) and summarized the curve as the average precision (*B*). We then computed a precision-recall curve for each individual feature using untransformed scores. The numbers of pathogenic (p) and benign (b) variants investigated are shown at the *bottom left* of each panel in *B*. For all epilepsy variants, PathoPredictor performed significantly better than any single feature ($P < 10^{-4}$), and PathoPredictor only failed to be significantly better in six of the 24 total feature comparisons (CCR, VEST, and missense depletion for RASopathies, CCR for dominant epilepsy genes, and missense depletion and MTR for cardiomyopathy).

Discussion

We have shown that the efficacy of variant pathogenicity prediction varies by disease, whereby each disease dictates a unique combination of classifier features. We have also presented PathoPredictor, a new missense variant pathogenicity predictor trained with variants from clinical sequencing results to produce pathogenicity scores from disease-specific combinations of regional constraint and variant features. PathoPredictor achieves an average precision greater than its components: CCR, FATHMM, missense depletion, missense badness, MTR, Pfam domain status, and VEST. FATHMM, Pfam domain status, and VEST capture regional constraint by using domains and protein families, whereas CCR, missense depletion, missense badness, and MTR locate genomic regions with less natural population variants than expected by null models of variation.

The evaluation of PathoPredictor and other variant classification tools is limited by available data describing pathogenic and benign variants. Ideally, these data would come from unbiased functional, mechanistic, tissue-based studies. Because these data sets do not exist in large quantities, we chose to use ClinVar and established variant interpretation protocols to determine a ground truth for variant pathogenicity. The performance of PathoPredictor is dependent on the quality of these annotations, and additional functional studies are needed to construct better databases for the training and evaluation of pathogenicity classifiers.

CCR was determined to be the most useful feature for classification, replicating results from Havrilla et al. (2019). Consistent with a recent survey of pathogenicity predictor performance using ClinVar variants (Ghosh et al. 2017), we found that VEST outperformed FATHMM for ClinVar variants. Missense depletion and badness were consistently the worst performing classification scores. The differing performance of these tools by disease panel demonstrates the utility of constructing PathoPredictor as a disease-specific combination of tools.

To construct PathoPredictor, we introduced a unique variant data set derived from clinical panel sequencing

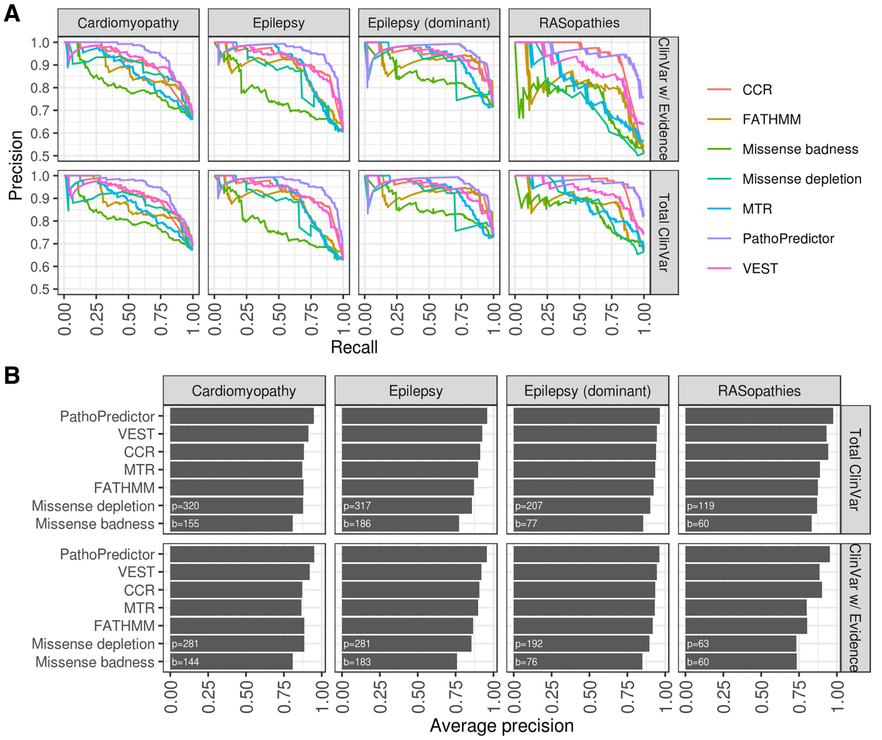

Figure 4. Disease-specific classifier performance using disease panel data for training and ClinVar data for testing. For each disease panel, we applied the hold-one-gene-out models from Figure 3 to ClinVar variants from the held-out gene to obtain pathogenicity prediction scores. We compared PathoPredictor to each feature using a precision-recall curves (*A*) and average precisions summarizing each curve (*B*). We used either all ClinVar variants (Total ClinVar) or ClinVar variants with a review status that included at least one submitter or an expert panel (ClinVar w/Evidence). The numbers of pathogenic (p) and benign (b) variants investigated are shown at the *bottom left* of each panel in *B*. PathoPredictor performs significantly better than any single feature when examining all ClinVar variants ($P < 0.03$).

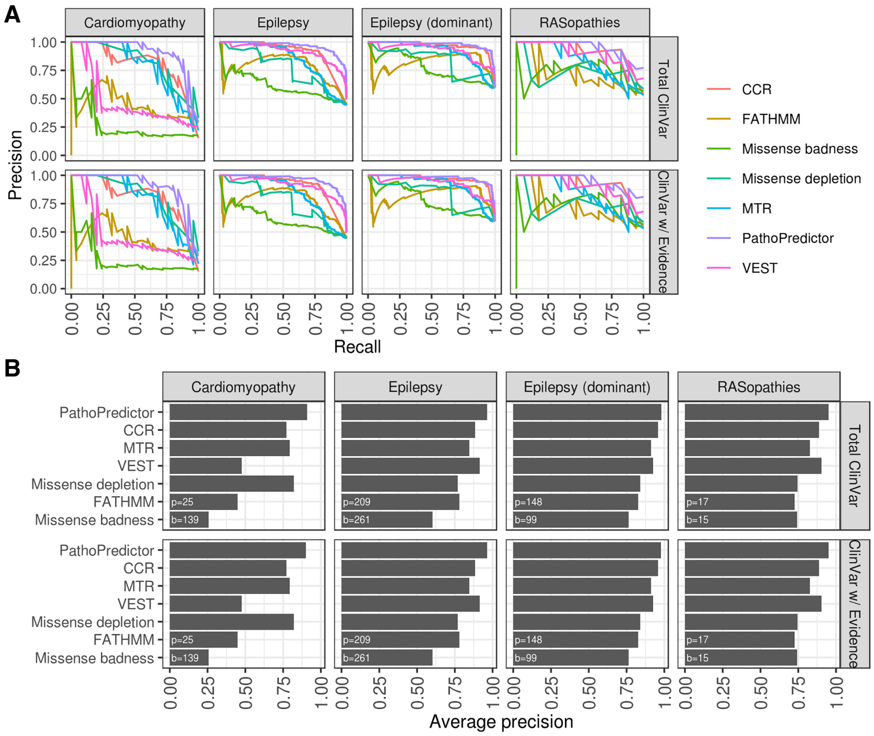

Figure 5. Disease-specific classifier performance using ClinVar data for training and disease panel data for testing. For each disease panel, we collected ClinVar variants in panel genes, using either all ClinVar variants (Total ClinVar) or reviewed ClinVar variants (ClinVar w/Evidence). PathoPredictor training and evaluation for each disease panel proceeded with a hold-one-gene approach. Disease panel variants from the gene of interest were used for evaluation, and ClinVar variants from all remaining disease panel genes were used for training. Using the held-out gene variant prediction scores, we computed a precision-recall curve (*A*) and summarized the curve with its average precision (*B*). We then computed a precision-recall curve for each individual feature using untransformed scores. The numbers of pathogenic (p) and benign (b) variants investigated are shown at the *bottom left* of each panel in *B*. PathoPredictor performed better than each of its six features ($P < 0.05$), except for missense depletion for cardiomyopathy panel variants, CCR, missense badness, and VEST for RASopathy panel variants, and CCR for dominant epilepsy panel variants.

results for cardiomyopathy, epilepsy, and RASopathy patients. A benefit of this data set compared to ClinVar is that the variants are labeled and obtained in a more homogeneous way, which helps remove data biases. Furthermore, the variant labels followed clinical interpretation standards similar to the ACMG/AMP guidelines, making the data set more similar to real-world clinical use cases. The clinically classified (pathogenic and benign) variants incorporated several pieces of evidence, mainly segregation, variant effect, functional, and allele frequency data, with limited reliance on computer predictions (or none), thus ensuring that no biases exist toward any one prediction tool. To avoid any further biases in our training and test data sets, we removed all variants previously used to train any of the component features. However, this significantly reduced the number of variants to optimize and evaluate PathoPreditor. Further testing and optimization, with larger clinically curated variant data sets, is required to confirm PathoPredictor's superior performance, and its utility in a clinical setting. An additional limitation of PathoPredictor, and other pathogenicity scores mentioned here, is that they are trained and evaluated using missense variants, ignoring synonymous variants that may impact splicing.

We demonstrated the utility of PathoPredictor using missense variants from ClinVar and a variant set of de novo variants previously used to compare REVEL and CCR. PathoPredictor performed significantly better than its constituent features when evaluated with ClinVar. However, a recent study of ClinVar variants concluded that although ClinVar has improved over time, it contains incorrect pathogenic labels for some endocrine tumor syndrome variant labels (Toledo and NGS in PPGL (NGSnPPGL) Study Group 2018), as an example. Although this ClinVar problem could affect our results, we also found that PathoPredictor had a significantly higher average precision than REVEL when testing with the de novo variants, which is consistent with CCR's improvement over REVEL using this same data set (Havrilla et al. 2019).

In conclusion, we recommend using PathoPredictor scores to predict missense variant pathogenicity for cardiomyopathy, epilepsy, and RASopathies. Predictions for all possible missense variants for disease panel genes are located in Supplemental Table S2.

Methods

Classifier

We used Python's scikit-learn machine learning library to train a logistic regression model to predict the pathogenicity of missense variants from clinical panels. Variants were classified by two well-known clinical laboratories, GeneDx and the Laboratory for Molecular Medicine (Harvard Medical School), using variant interpretation protocols that are well within the most recent 2015 ACMG/AMP guidelines. Pathogenic and likely pathogenic variants were assigned values of one, and benign and likely benign variants were assigned values of zero. During training, we used L2 regularization with a regularization strength of one. Our model included six features corresponding to measures of pathogenicity, one Pfam domain indicator, and all pairwise combinations of features. All model terms were standardized by removing the mean and scaling by the standard deviation with scikit-learn.

Pathogenicity scores as classifier features

We used seven features in our classifier: CCR, FATHMM, missense depletion, missense badness, MTR, VEST, and Pfam protein domains. FATHMM and VEST can provide multiple scores for one variant, depending on isoforms. VEST scores were taken as the minimum VEST v3.0 score provided by dbNSFP v2.9 (Liu et al. 2013). FATHMM scores were taken as the negative minimum FATHMM v2.3 score provided by dbNSFP v2.9. FATHMM scores were negated so that their interpretation would match the other features. CCR scores were taken as the CCR percentile (ccr_pct) from the CCR BED file v1.20171112 (Havrilla et al. 2019). Missense depletion and badness scores were taken from the constraint MPC VCF file v2 as obs_exp and mis_badness, respectively (Samocha et al. 2017). Missense depletion was negated so that its

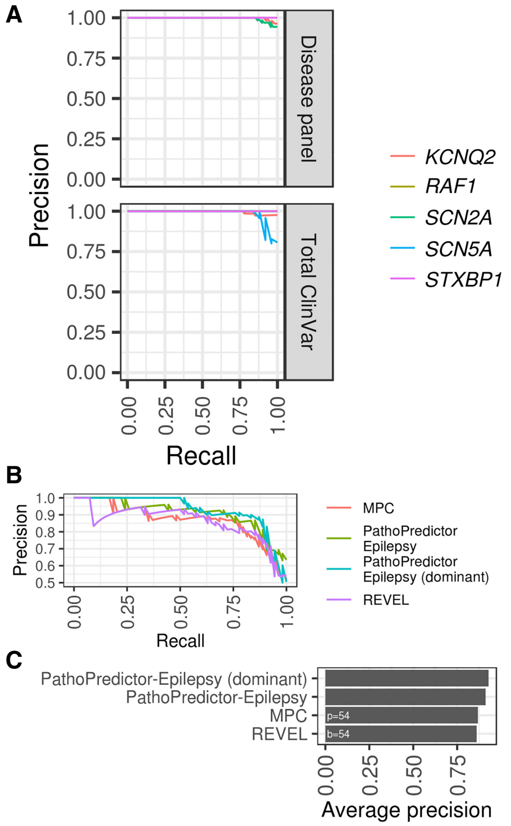

Figure 6. PathoPredictor performance. (*A*) Precision-recall curves are shown for select genes evaluated during cross-validation with the disease panel data set and tested with ClinVar variants. The curve for *RAF1* closely follows and is obscured by that of *SCN2A*. For *KCNQ2*, PathoPredictor had an accuracy of 95% for panel variants and 96% for ClinVar variants. (*B*) PathoPredictor epilepsy-specific classifiers were compared to REVEL and MPC. De novo missense variants in epilepsy panel genes were used as pathogenic variants. Epilepsy panel gene missense variants from unaffected siblings of autism spectrum disorder patients were used as benign variants. PathoPredictor was trained as in Figure 4, but only utilizing the full and dominant epilepsy data sets. Variants were filtered using the same methods applied to ClinVar variants, and additional filters were applied to remove training data for MPC. (*C*) We summarized each scoring metric's precision-recall curve as the average precision. Both PathoPredictor classifiers achieved a greater average precision than REVEL ($P < 0.05$), and the dominant epilepsy classifier performed better than MPC ($P < 0.05$). (p) pathogenic; (b) benign.

interpretation would match the other features. Traynelis et al. (2017) provided chromosome-specific tab delimited text files containing MTR scores and associated metrics for single genomic positions. We extracted MTR scores, negated them so that their interpretation would match the other features, and constructed BED files in which each line corresponded to a region of consecutive positions with identical MTR scores. A BED file of Pfam domain locations was downloaded from the UCSC Genome Browser. We assigned each variant a Pfam score of one if its position fell within a domain, and zero otherwise. Note that we omitted this simple Pfam domain feature from figures comparing feature classification performance because we did not expect it to perform well by itself. Feature values were assigned to variants as described below in the variant annotation and filtering pipeline.

For each disease data set, we used Python's scikit-learn library to standardize each feature by removing its mean and scaling by its standard deviation. Disease panel, ClinVar, and de novo variants for the same disease were processed together so that their features would be on the same scale.

Variant sets

We used three missense variant sources in this study: disease panels and ClinVar for model training and validation (using unique variant sets), and neurodevelopmental patients for comparing PathPredicor to MPC and REVEL.

GeneDx provided clinical sequencing panel results for epilepsy, and the Laboratory for Molecular Medicine provided their clinically curated data for cardiomyopathy and RASopathies. The number of patients investigated differed by gene. The maximum number of patients observed was 5466 for cardiomyopathy, 8583 for epilepsy, and 3022 for RASopathies. No gene was shared between the three data sets. Variants were provided in Human Genome Variation Society (HGVS) c. notation (Dunnen et al. 2016), and were converted to VCF files of hg19-based variants using Mutalyzer (Wildeman et al. 2008) and custom scripts (Supplemental Code). To construct variant sets for our classifier, we discarded variants of uncertain significance. We formed a benign set of variants using "benign" and "likely benign" variants. Similarly, our pathogenic variant set consisted of "pathogenic" and "likely pathogenic" variants. These labeled variants were used for training disease-specific classifiers (see below). Diseases panel variants and labels were deposited into Supplemental Table S1.

Variants from ClinVar were chosen as a validation set. We restricted ClinVar genes to those found in the disease panels and removed any ClinVar genomic position found in the disease panels, producing an independent variant set. The hg19 ClinVar VCF file was downloaded on February 25, 2018, and limited to unambiguously pathogenic or likely pathogenic and benign or likely benign variants with no conflicts according to CLINSIG (Landrum et al. 2016). We considered ClinVar variants with any review status as one test set and consulted CLNREVSTAT (Landrum et al. 2016) to produce a second ClinVar test variant set restricted to reviewed variants.

As in the CCR paper (Havrilla et al. 2019), we compared PathoPredictor, REVEL, and MPC using de novo missense variants from 5620 neurodevelopmental disorder patients and 2078 unaffected siblings of autism spectrum disorder patients (Samocha et al. 2017; Havrilla et al. 2019). De novo variants from patients were considered pathogenic, and de novo variants from unaffected siblings were considered benign. HGVS formatted variants were uploaded to VariantValidator (Freeman et al. 2018), and a VCF file was constructed from the results. This file was normalized with vt (Tan et al. 2015). To avoid evaluating with any tool's training data, we removed disease panel variants, ClinVar variants, and benign variants present in >1% of ExAC (MPC's benign training data).

Disease-specific classifier evaluation

We compared estimated pathogenicity probabilities produced by our trained models with each pathogenicity score used as a model feature via precision-recall curves and average precision, as implemented in scikit-learn. Precision-recall curves and average precision are useful here because of the possibility of imbalances between pathogenic and benign variant counts. Average precision is an approximation of the area under the method's precision-recall curve. We ran three experiments with disease panel and ClinVar variants to evaluate the performance of PathoPredictor. First, we used cross-validation with disease panel variants. Second, we trained a model with disease panel variants and validated it with ClinVar variants. Third, we trained a model with ClinVar variants and validated it with disease panel variants.

Comparisons were conducted in a leave-one-gene-out manner. We iterated over all genes in a disease, holding out the gene of interest and training a model using variants from all remaining genes. This model was applied to variants from the gene of interest, ensuring that a given gene was never used for training and validation. ClinVar variant data sets were restricted to variants not found in the disease panel results. Precision-recall curves and average precision scores were made for each pathogenicity score by aggregating the results from each gene evaluation. The DeLong test as implemented in R's pROC package (Robin et al. 2011) was used to compare areas under receiver operating characteristics curves produced by predictors. We used this test to gauge the significance of differences between classifiers.

Comparing PathoPredictor with MPC and REVEL

We applied our epilepsy variant trained PathoPredictor (using all or dominant epilepsy genes) to de novo missense variants. For the evaluation of PathoPredictor, MPC, and REVEL, we used 54 pathogenic missense variants located in the epilepsy panel genes. Limiting the benign missense data set to epilepsy genes produced only six benign variants for evaluation, so we randomly selected 48 variants from the full benign missense data set of 969 variants not located in epilepsy genes so that the pathogenic and benign evaluation sets would have the same size. PathoPredictor, MPC, and REVEL scores were compared using precision-recall curves, average precision, and the DeLong test.

Variant annotation and filtering pipeline

Our pipeline began with VCF files containing 6382 de novo, 345,849 ClinVar, and 7840 disease panel variants labeled as benign, pathogenic, or variant of unknown clinical significance. SnpEff v4.3.1T (Cingolani et al. 2012) was used to determine variant effects in GRCh37.75. SnpSift v4.3 (Cingolani et al. 2012) was used to annotate variants with allele frequencies from the ESP, FATHMM scores, and VEST scores from dbNSFP. We annotated variants with values from BED (CCR and Pfam) and VCF (missense badness and depletion and calculated ESP frequencies from ESP6500SI-V2) files using vcfanno v0.2.8 (Pedersen et al. 2016). The ESP frequencies are needed next when removing the training data used for VEST.

We next removed variants that had been used to train FATHMM (~49,500 disease variants and ~37,000 putatively neutral variants) or VEST (~45,000 disease variants and ~45,000 putatively neutral variants), ensuring that these features would not have an advantage when comparing pathogenicity scores and that our validation data sets would not overlap with any variants used for training. Both FATHMM and VEST were trained with damaging mutations from HGMD, but they differed in their choice of neutral missense variant set. FATHMM was trained with neutral variants from UniProt (The UniProt Consortium 2017), and VEST was trained with missense variants from ESP achieving a population frequency of 1% or higher.

We then removed variants found in the set of 154,257 DM (damaging mutation) in HGMD Professional 2016.1. To address frequent ESP variants, we took the variant frequency as the maximum of dbNSFP fields ESP6500_EA_AF, ESP6500_AA_AF, and the total ESP allele frequency determined using vcfanno. We discarded variants (68 for ClinVar and 59 disease panel) when this maximum value reached at least 0.01. To remove neutral UniProt variants, we used "Polymorphism" annotations to build a list of neutral codons relative to hg38. Polymorphism annotations were downloaded (www.uniprot.org/docs/humsavar.txt) and joined with hg38 codon coordinates from UniProt. Both were downloaded on April 5, 2018. We used liftOver (Hinrichs et al. 2006) to convert these to hg19 and removed any variant found in any of 921,722 neutral codons.

Final variant sets were taken as missense variants with CCR and missense depletion and MTR scores to avoid missing data issues. After applying all the aforementioned filters, 666 disease panel, 1159 ClinVar, and 108 de novo missense variants were used in this study.

Software availability

Source code for this manuscript is available at https://github.com/samesense/pathopredictor and included as Supplemental Code. A docker image for running PathoPredictor is available at https://hub.docker.com/r/samesense/pathopredictor/. Diseases panel variants and labels were deposited into Supplemental Table S1. Predictions for all possible missense variants for disease panel genes are located in Supplemental Table S2.

Competing interest statement

A.L., D.A.M., M.L., and M.S. work fully or partially in fee-for-service genetic laboratories.

Acknowledgments

We thank all members of GeneDx and the Laboratory for Molecular Medicine (LMM) who manually curated and classified the disease variants used in our study.

References

Amr SS, Al Turki SH, Lebo M, Sarmady M, Rehm HL, Abou Tayoun AN. 2017. Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation. *Genet Med* **19:** 496–504. doi:10.1038/gim.2016.134

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 (**Suppl 3**):** S3. doi:10.1186/1471-2164-14-S3-S3

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso*-2; *iso*-3. *Fly (Austin)* **6:** 80–92. doi:10.4161/fly.19695

Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE. 2016. HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat* **37:** 564–569. doi:10.1002/humu.22981

Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* **32:** 557–563. doi:10.1002/humu.21438

Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. 2018. VariantValidator: accurate validation, mapping, and formatting of sequence variation descriptions. *Hum Mutat* **39:** 61–68. doi:10.1002/humu.23348

Ghosh R, Oak N, Plon SE. 2017. Evaluation of *in silico* algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol* **18:** 225. doi:10.1186/s13059-017-1353-5

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313:** 903–919. doi:10.1006/jmbi.2001.5080

Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. 2019. A map of constrained coding regions in the human genome. *Nat Genet* **51:** 88–95. doi:10.1038/s41588-018-0294-6

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34** (Database issue): D590–D598. doi:10.1093/nar/gkj144

Homburger JR, Green EM, Caleshu C, Sunitha MS, Taylor RE, Ruppel KM, Metpally RP, Colan SD, Michels M, Day SM, et al. 2016. Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation. *Proc Natl Acad Sci* **113:** 6701–6706. doi:10.1073/pnas.1606950113

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, et al. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* **99:** 877–885. doi:10.1016/j.ajhg.2016.08.016

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44:** D862–D868. doi:10.1093/nar/gkv1222

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536:** 285–291. doi:10.1038/nature19057

Liu X, Jian X, Boerwinkle E. 2013. dbNSFP v2.0: a database of human nonsynonymous SNVs and their functional predictions and annotations. *Hum Mutat* **34:** E2393–E2402. doi:10.1002/humu.22376

Pedersen BS, Layer RM, Quinlan AR. 2016. *Vcfanno*: fast, flexible annotation of genetic variants. *Genome Biol* **17:** 118. doi:10.1186/s13059-016-0973-5

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. 2015. 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* **17:** 405–423. doi:10.1038/gim.2015.30

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M, et al. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12:** 77. doi:10.1186/1471-2105-12-77

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, et al. 2014. A framework for the interpretation of *de novo* mutation in human disease. *Nat Genet* **46:** 944–950. doi:10.1038/ng.3050

Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ. 2017. Regional missense constraint improves variant deleteriousness prediction. bioRxiv doi:10.1101/148353

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34:** 57–65. doi:10.1002/humu.22225

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. *Genome Med* **1:** 13. doi:10.1186/gm13

Tan A, Abecasis GR, Kang HM. 2015. Unified representation of genetic variants. *Bioinformatics* **31:** 2202–2204. doi:10.1093/bioinformatics/btv112

Tavtigian SV, Byrnes GB, Goldgar DE, Thomas A. 2008. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Hum Mutat* **29:** 1342–1354. doi:10.1002/humu.20896

Toledo RA, NGS in PPGL (NGSnPPGL) Study Group. 2018. Inflated pathogenic variant profiles in the ClinVar database. *Nat Rev Endocrinol* **14:** 387–389. doi:10.1038/s41574-018-0034-0

Traynelis J, Silk M, Wang Q, Berkovic SF, Liu L, Ascher DB, Balding DJ, Petrovski S. 2017. Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res* **27:** 1715–1729. doi:10.1101/gr.226589.117

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45:** D158–D169. doi:10.1093/nar/gkw1099

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* **29:** 6–13. doi:10.1002/humu.20654

Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. 2011. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27:** 2147–2148. doi:10.1093/bioinformatics/btr357

Received June 23, 2018; accepted in revised form May 24, 2019.