



## Research and Applications

# Evaluating dimensionality reduction of comorbidities for predictive modeling in individuals with neurofibromatosis type 1

Aditi Gupta, PhD<sup>\*1</sup>, Ethan Hillis, MS<sup>1</sup>, Inez Y. Oh , PhD<sup>1</sup>, Stephanie M. Morris, MD<sup>2</sup>, Zach Abrams, PhD<sup>1</sup>, Randi E. Foraker, PhD, MA<sup>1</sup>, David H. Gutmann, MD, PhD<sup>3</sup>, Philip R. O. Payne , PhD<sup>1</sup>

<sup>1</sup>Institute for Informatics, Data Science and Biostatistics, Washington University, Saint Louis, MO 63110, United States, <sup>2</sup>Center for Autism Services, Science, and Innovation (CASSI), Kennedy Krieger Institute, Baltimore, MD 21205, United States, <sup>3</sup>Department of Neurology, School of Medicine, Washington University, Saint Louis, MO 63110, United States

<sup>\*</sup>Corresponding author: Aditi Gupta, PhD, Institute for Informatics, Data Science and Biostatistics, Washington University, 660 South Euclid Avenue, Campus Box 8132, Saint Louis, MO 63110, United States (agupta24@wustl.edu)

A. Gupta and E. Hillis contributed equally and are considered co-first authors of this work.

## Abstract

**Objective:** Dimensionality reduction techniques aim to enhance the performance of machine learning (ML) models by reducing noise and mitigating overfitting. We sought to compare the effect of different dimensionality reduction methods for comorbidity features extracted from electronic health records (EHRs) on the performance of ML models for predicting the development of various sub-phenotypes in children with Neurofibromatosis type 1 (NF1).

**Materials and Methods:** EHR-derived data from pediatric subjects with a confirmed clinical diagnosis of NF1 were used to create 10 unique comorbidities code-derived feature sets by incorporating dimensionality reduction techniques using raw International Classification of Diseases codes, Clinical Classifications Software Refined, and Phecode mapping schemes. We compared the performance of logistic regression, XGBoost, and random forest models utilizing each feature set.

**Results:** XGBoost-based predictive models were most successful at predicting NF1 sub-phenotypes. Overall, features based on domain knowledge-informed mapping schema performed better than unsupervised feature reduction methods. High-level features exhibited the worst performance across models and outcomes, suggesting excessive information loss with over-aggregation of features.

**Discussion:** Model performance is significantly impacted by dimensionality reduction techniques and varies by specific ML algorithm and outcome being predicted. Automated methods using existing knowledge and ontology databases can effectively aggregate features extracted from EHRs.

**Conclusion:** Dimensionality reduction through feature aggregation can enhance the performance of ML models, particularly in high-dimensional datasets with small sample sizes, commonly found in EHRs health applications. However, if not carefully optimized, it can lead to information loss and data oversimplification, potentially adversely affecting model performance.

## Lay Summary

Dimensionality reduction, a technique used to simplify data by reducing noise and overfitting, plays a key role in enhancing the performance of machine learning (ML) models. This study assessed various dimensionality reduction methods applied to comorbidity features extracted from the electronic health records (EHRs) of children with Neurofibromatosis type 1 (NF1). Due to extreme heterogeneity in the clinical sub-phenotypes arising in people with NF1, it is difficult to predict who will develop one or more of the many NF1-associated clinical sub-phenotypes. Using the reduced feature sets derived from diagnostic codes, 3 ML models were employed to predict NF1 sub-phenotypes such as optic pathway glioma, attention-deficit hyperactivity disorder, and scoliosis. The study demonstrated that model performance is significantly impacted by the choice of dimensionality reduction technique and varies depending on the specific ML algorithm and the predicted outcome. Automated methods utilizing existing knowledge and ontology databases can effectively aggregate features derived from EHRs. Feature aggregation through dimensionality reduction can significantly boost ML model performance, particularly in high-dimensional datasets with small sample sizes, which are common in EHR-based health applications. However, if not carefully optimized, dimensionality reduction can lead to information loss and data oversimplification, potentially negatively affecting model performance.

**Key words:** clinical research informatics; electronic health records; predictive modeling; neurofibromatosis type 1.

## Background and significance

High-dimensional data that describe the health status of a particular individual are available from various sources, including

electronic health records (EHRs), imaging, wearables, patient reported data, and genomic datasets. These sources provide informative data for artificial intelligence (AI)-based approaches

Received: October 19, 2024; Revised: December 16, 2024; Editorial Decision: December 21, 2024; Accepted: December 24, 2024

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

to predicting disease diagnosis and progression. Integrating and harmonizing health data from various sources theoretically improves quality and accuracy, but it also increases complexity and redundancy of the data. This problem is often referred to as “the curse of dimensionality,” and is particularly challenging when analyzing healthcare information.<sup>1</sup> To develop performant and potentially generalizable AI-based models for health applications with small sample sizes, there is a pressing need to create methods that can select important and prognostic feature sets from high-dimensional raw input datasets.

EHRs provide clinically relevant, longitudinal information about the medical history of any given person, such as demographics, laboratory results, vital signs, procedures, medications, and comorbidities. Feature reduction and aggregation methods allow raw EHR data to be transformed into standardized high-level features, which can then be used to develop predictive models. Comorbidities, often recorded as diagnoses or medical history in the EHR, have been previously shown to be informative for predictive models.<sup>2–4</sup> International Classification of Diseases, Tenth Revision, Clinical Modification (ICD10-CM) diagnosis codes can be challenging to use in their raw form to represent comorbidities due to their high-dimensional data values, creating sparse feature sets. They have a hierarchical structure consisting of chapters, blocks of categories, individual categories, subcategories, and subdivisions that organize diagnoses by type of medical condition, anatomical location, and severity.<sup>5</sup> Approximately 69 000 unique ICD10-CM codes and various coding schemas have been proposed to categorize and combine them into groups.<sup>6</sup> Phecodes are a popular mapping method for grouping codes into higher-level categories, reflecting distinct diseases or traits.<sup>7</sup> Another commonly used database for aggregating codes into clinically meaningful comorbidity categories is the Clinical Classifications Software Refined (CCSR), developed as part of the Healthcare Cost and Utilization Project (HCUP) sponsored by the Agency for Healthcare Research and Quality (AHRQ).<sup>8</sup>

While diagnostic mapping terminologies can improve the performance of predictive models, relative performance varies given the particular machine learning (ML) algorithm employed and the prediction outcome of interest.<sup>9</sup> To develop clinical predictive models, unsupervised ML techniques for dimensionality reduction have also been used. For example, one study found that dimensionality reduction models, such as principal component analysis (PCA), demonstrated a slight improvement relative to stepwise logistic regression for predicting 30-day major adverse cardiac events in individuals presenting to the emergency department with chest pain.<sup>10</sup> However, to our knowledge, there has not been a comprehensive comparison of feature-aggregating mapping schema and unsupervised feature reduction methods to reduce the dimensionality of comorbidity information relevant to the success of a predictive model.

For rare diseases or other conditions with small data cohorts, “the curse of dimensionality” is particularly exacerbated, and it is unlikely for any one hospital or clinic to have more than a handful of individuals with the condition of interest. Herein, we employ Neurofibromatosis type 1 (NF1), a rare autosomal dominant syndrome affecting 1 in 2800 people worldwide, as a test case.<sup>11,12</sup> Due to extreme heterogeneity in the clinical sub-phenotypes arising in people with NF1, even in family members with the same germline *NF1* mutation, it is difficult to predict who will develop one or

more of the many NF1-associated clinical sub-phenotypes. A phenotype or sub-phenotype is a clinical entity defined by observable characteristics that is produced by an interaction between the genotype and the environment.<sup>13</sup> Children with NF1 are at an increased risk of developing a variety of sub-phenotypes, including optic pathway glioma (OPG),<sup>14</sup> attention-deficit/hyperactivity disorder (ADHD),<sup>15</sup> and scoliosis.<sup>16</sup> The ability to predict the risk of developing these clinical sub-phenotypes would enable improved monitoring, early intervention, and enhanced outcomes and quality of life for affected children and their families. In this study, we compare the performance of predictive models using different dimensionality reduction methods for comorbid features extracted exclusively from the EHR to predict the development of OPG, ADHD, and scoliosis in children with NF1. Here, we compare the performance of models using reduced comorbidity feature sets based on raw ICD-10-CM codes, CCS, and Phecode mapping schemes, as well as clustering methods.

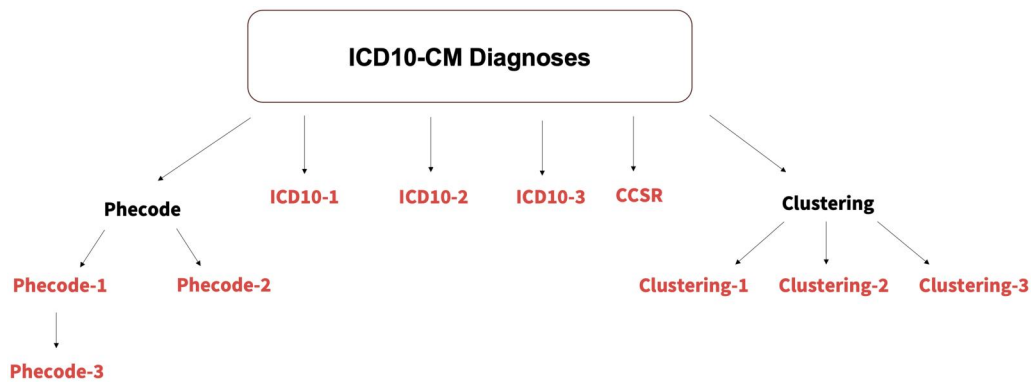
## Materials and methods

With the approval of the Washington University Institutional Review Board (protocol #201706112), EHR data were extracted from the Research Data Core at Washington University School of Medicine, a data lake containing retrospective clinical records from Washington University Medicine and BJC Healthcare, including those from Barnes-Jewish Hospital and St Louis Children’s Hospital. All subjects were younger than 18 and had a confirmed clinical diagnosis of NF1 (ICD10-CM code Q85.01). The extracted data included features from structured data, comprising all diagnosed ICD10-CM codes within the cohort and demographic information. We created 10 datasets for comparison, keeping sex and race constant across each while varying the ICD-10-CM-derived features by using different feature reduction and aggregation methods to modify the diagnoses and group them into their clinical context-based feature sets (Figure 1). To ensure that the results were not phenotype-specific, three different sub-phenotypes (OPG, ADHD, and scoliosis) were chosen for analysis. Importantly, while measures have been taken to ensure that the results are as unbiased and accurate as possible, the resulting predictive models and results were not evaluated for clinical accuracy.

## Data inclusion for predictive modeling

Individuals diagnosed with OPG, ADHD, or scoliosis were identified using ICD10-CM codes. Physician-experts in the field of NF1 (S.M.M. and D.H.G.) curated a list of ICD10-CM codes for each of the 3 conditions (Table S1).

Unique feature sets were created for training and evaluating the predictive models for OPG, ADHD, and scoliosis. For each of these conditions, individuals who were not diagnosed with the condition within the timeframe of the extracted data had their full available diagnosis history used to determine the presence or absence of each of the ICD10-CM diagnoses present in the study cohort. Individuals diagnosed with a condition had only their diagnosis history included in training the prediction model until their diagnosis was rendered. The datasets were constructed in this way to prevent leakage of the outcome variables into the dataset used for prediction.



**Figure 1.** ICD10-CM feature aggregation and reduction workflow.

### Diagnosis feature aggregation and reduction

Various dimensionality reduction methods created 10 total feature sets. Seven were built using methods including aggregating features based on ICD10-CM coding hierarchies, mappings that group together hierarchical ICD10-CM codes, and prevalence-based filtering to keep the most common comorbidities. The remaining 3 feature sets were generated using unsupervised clustering and PCA techniques. (Table 1; further described below).

### Feature sets using ICD-based methods

The base dataset (ICD10-1) included all 2982 unique ICD10-CM codes within the cohort represented as binary variables. We used Phecode and CCSR, aggregation and reduction methods that group ICD-10 codes into standardized, clinically meaningful concepts<sup>5–8</sup> (Table 1). Phecode and CCSR map codes into concepts, or features, by aggregating codes that are clinically similar. Each has different hierarchies that aggregate and reduce the feature space in different ways.

### Feature sets using clustering and PCA methods

Each of these feature sets was created using a different method, namely k-means clustering,<sup>17</sup> hierarchical clustering, and PCA<sup>18</sup> (Table 1).

The matrix for clustering was a square matrix [2982\*2982], holding the counts of every pairwise set of ICD10-CM codes in our cohort. In this matrix, cell  $C_{i,j}$  contains the number of individuals that have cooccurring diagnoses of both  $i$  and  $j$  ICD10-CM codes.

For k-means and hierarchical clustering, the elbow method was used to determine the optimal k clusters, where each cluster contains diagnoses that frequently co-occur in individuals in our cohort. Yellowbrick, a ML visualization library, was used for this purpose.<sup>19</sup> For k-means clustering, we used the Lloyd algorithm with K-Means++ centroid initialization, which samples and assigns initial centroids based on the maximum squared distance from each other. Hierarchical clustering used Euclidean distance and Ward linkage. For PCA, we set a threshold of 95% variance explained to select the number of principal components to use ( $N = 98$ ).

### ML models

The effect of the different feature sets on predictive modeling was evaluated using 3 different algorithms: logistic regression,<sup>20</sup> random forest,<sup>21</sup> and XGBoost.<sup>22</sup> Model selection and hyperparameter tuning were done using nested cross-

validation and randomized search with 100 parameter space samples for each inner fold. From this, 5 parameter sets were generated—one for each top-performing model from each outer fold was returned. The highest performing parameter set according to F1-score was selected as the final parameter set and evaluated by averaging performance over a series of 5-fold cross-validation iterations. The splits were kept the same across models and datasets to ensure an unbiased comparison of performances. The evaluation metrics were area under the receiver operating characteristic curve (AUROC), average precision, and F1 score.

Additionally, when mapping to different feature sets, specific ICD10-CM codes were filtered out to prevent leakage of the outcome variable into the set of predictor variables. All feature sets were created by starting with the base ICD10-CM dataset (ICD10-1), removing the ICD10-CM codes corresponding specifically to the outcome variable, then mapping the remaining codes to each different feature set. For example, a feature in Phecode Expanded named “Cancer of brain and nervous system” encompassed diagnoses related to brain and nervous system cancers, including several codes used to identify OPG subjects in our cohort. To keep this feature in the model, those codes used to identify subjects with OPG in our data were removed before mapping so they did not leak into the dataset. Similarly, for the ICD10-1 and ICD10-2 OPG, ADHD, and scoliosis datasets, the identifying codes for these sub-phenotypes were removed from their respective feature space.

### Statistical analysis

Statistical testing was conducted to compare the performance between feature sets. Friedman’s test, a nonparametric test of repeated measures, is commonly used for comparing multiple classifiers over multiple datasets.<sup>23</sup> More specifically, Friedman’s test is appropriate in situations where the independent variable is binary, there are evaluation metrics from multiple testing sets, and more than 2 models are being compared. These conditions apply to our study. The post hoc test selected was the Conover test, which is commonly used to perform pairwise comparisons following Friedman’s test.<sup>24</sup>

Initially, we used Friedman’s test to compare AUROC values from each feature set across different conditions and models. Subsequently, the Conover test, with Bonferroni correction for multiple comparisons, was performed to identify statistically significant pairwise differences. A significance threshold of 0.05 was used for all tests.

**Table 1.** Feature set names, descriptions, and *N* features.

Feature set acronym	Feature set	Description	Features ( <i>n</i> )
ICD10-1	ICD10-CM Base	Base feature set with each ICD10-CM diagnosis code in the cohort represented as a binary feature	2982
ICD10-2	ICD10-CM Selected	Subset of features from “ICD10-CM Base” based on a individual prevalence greater than 1%	477
ICD10-3	ICD10-CM First 3	Maps ICD10-CM code features from “ICD10-CM Base” by using only the first 3 digits	786
Phecode-1	Phecode expanded	Map all ICD10-CM code features from “ICD10-CM Base” to lowest level Phecode categories	861
Phecode-2	Phecode expanded selected	Subset of features from “Phecode Expanded” based on an individual prevalence greater than 1%	299
Phecode-3	Phecode exclusive	Map all ICD10-CM code features from “ICD10-CM Base” to highest level Phecode categories	22
CCSR	CCSR	Categorize ICD10-CM code features from “ICD10-CM Base” using CCS standard mapping	309
Clustered-1	K-means clustered	K-means clustering-based feature reduction using Lloyd algorithm and “K-means++” centroid initialization	252
Clustered-2	Hierarchical clustered	Hierarchical clustering-based feature reduction using Euclidean distance and Ward linkage	232
Clustered-3	PCA	PCA-based feature reduction	98

Abbreviations: ICD = International Classification of Diseases; ICD10-CM = International Classification of Diseases, Tenth Revision, Clinical Modification; CCSR = Clinical Classifications Software Refined; PCA = Principal Component Analysis.

### Feature importance analysis

We analyzed the most important features of the various feature sets to determine, for each outcome variable, whether the most critical features of the feature space are similar across feature sets. This similarity was quantified by the number of features shared by the top 10 most important features of each feature set as identified by XGBoost.

However, it is challenging to perform these comparisons for feature sets derived by grouping ICD10-CM codes (including everything except for ICD10-CM Base [ICD10-1] and ICD10-CM Selected [ICD10-2]). For this reason, the top 10 features of ICD10-CM Base were chosen as the baseline against which the 9 other feature sets were compared instead of comparing all pairwise combinations of the ten feature sets.

For the feature sets that grouped ICD10-CM codes into higher-level features, a feature was counted as shared if the feature encompasses at least 1 code from the top 10 features of ICD10-1. For example, if 1 of ICD10-1's 10 most important features for a prediction was “A01.1,” and 1 of CCSR's 10 most important features was titled “Example Feature,” which was made up of codes “A01.1,” “B01.1,” and “C01.1.” Since this CCSR feature included “A01.1,” then this was counted as a shared feature. The total number of shared features between ICD10-1 and CCSR would be the number of times this occurs out of ten, since only the top ten most important features were analyzed.

For ICD10-2, the only other feature set with singular ICD10-CM codes as features and did not group, a feature was counted as shared if the same code from ICD10-1 was present in ICD10-2. In this manner, the total number of shared features between ICD10-1 and ICD10-2 was equivalent to the intersection of their top 10 most important features.

For PCA, feature overlap must be handled differently because the ICD10-CM codes that define the components are not mutually exclusive. As a result, feature overlap considers only the top 10 ICD10-CM codes by magnitude of coefficient of linear combination for each component. HoloViews, a data analysis and visualization library, was used to generate the shared feature plots.<sup>25</sup>

## Results

### Clinical population

Six hundred forty individuals met the criteria for inclusion (Table 2). Within the timeframe encompassed by the data, 99

individuals were identified to have OPG, 202 were identified to have ADHD, and 118 were identified to have scoliosis. There were fewer females than males diagnosed with OPG (46.5%) or ADHD (41.6%), but a higher percentage of females were diagnosed with scoliosis (51.7%). Of the individuals who developed OPG, ADHD, and scoliosis, 46.5%, 41.6%, and 51.7% were female, respectively.

### Feature set reduction and aggregation

There exist 82 991 ICD10-CM codes. The two Phecode map hierarchies were used: (1) Expanded (Phecode-1), which transforms all codes into 1755 features, and (2) Exclusive (Phecode-3), which maps all codes into 17 features (S1). The standard CCSR hierarchy transforms all codes into 490 features (Figure S1). Our cohort has 2982 unique ICD-10 codes and the number of features resulting from each of the various reduction and aggregation methods is shown in Table 1. The optimal number of clusters for k-means (Clustering-1) and hierarchical clustering (Clustering-2) were determined to be 250 and 230, shown in Figure 2. For PCA (Clustering-3), the number of components sufficient to explain 95% of the variance was 98 components for OPG, 98 components for ADHD, and 96 components for scoliosis. Table 1 displays the total number of features for each feature set, including the demographic variables.

### Model performance

Figure 3 illustrates the AUROC performance for various conditions (OPG, ADHD, and scoliosis) as a function of the number of features (log-scaled) used in the ML model. It shows that as the number of features increases, the AUROC performance generally tends to stabilize or improve for each condition. However, we also observe that over-aggregation of features results in information loss and decreased performance for all models.

Model performance results are shown in Table 3. The results for the 3 highest performing feature sets for each model and sub-phenotype are bolded. The metrics include area under the receiver operator curve (AUROC), and F1-score (F1).

Comparing AUROC ranges, the models we developed were most successful at predicting OPG (0.62-0.82), followed by scoliosis (0.60-0.72) and ADHD (0.59-0.69).



**Table 2.** Cohort characteristics.

Variable	Total ( <i>n</i> = 640 [100.0%])	OPG ( <i>n</i> = 99 [15.5%])	Non-OPG ( <i>n</i> = 541 [84.5%])	ADHD ( <i>n</i> = 202 [31.6%])	Non-ADHD ( <i>n</i> = 438 [68.4%])	Scoliosis ( <i>n</i> = 118 [18.4%])	Non-Scoliosis ( <i>n</i> = 522 [81.6%])
Age (years), median (IQR)	10.5 (6.0-15.6)	7.9 (4.6-10.7)	11.0 (6.5-15.9)	9.7 (6.7-12.4)	10.1 (5.4-15.6)	10.3 (5.9-15.3)	10.4 (7.4-13.2)
Sex (female), <i>n</i> (%)	316 (49.4%)	46 (46.5%)	270 (49.9%)	84 (41.6%)	232 (53.0%)	61 (51.7%)	255 (48.9%)
Race, <i>n</i> (%)							
White, <i>n</i> (%)	521 (81.4%)	91 (91.9%)	430 (79.5%)	165 (81.7%)	356 (81.3%)	106 (89.8%)	415 (79.5%)
Other/unknown, <i>n</i> (%)	12 (1.9%)	0 (0.0%)	12 (2.2%)	4 (2.0%)	8 (1.8%)	1 (0.9%)	11 (2.1%)
Black, <i>n</i> (%)	86 (13.4%)	7 (7.1%)	79 (14.6%)	29 (14.3%)	57 (13.0%)	9 (7.6%)	77 (14.8%)
Asian, <i>n</i> (%)	21 (3.3%)	1 (1.0%)	20 (3.7%)	4 (2.0%)	17 (3.9%)	2 (1.7%)	19 (3.6%)

Abbreviations: OPG = Optic Pathway Glioma; ADHD = Attention-Deficit/Hyperactivity Disorder; IQR = Interquartile range.

XGBoost was the top-performing algorithm, producing the highest model performance for predicting OPG and ADHD and near the highest for scoliosis. Logistic regression had good predictive performance for OPG but poor performance for scoliosis.

The feature set CCSR performed best for predicting OPG, Phecode-2 for predicting ADHD, and ICD10-1 for predicting scoliosis.

The feature set Clustering-3 had the best performance for logistic regression, ICD10-2 had the best performance for random forest, and Phecode-2 had the best performance for XGBoost.

Overall, feature sets based on domain knowledge-informed mapping schema performed better than unsupervised feature reduction methods and raw ICD10-CM-based features. Some feature sets based on feature reduction methods outperformed mapping schema for certain combinations of sub-phenotype and algorithm, such as logistic regression with Clustering-3 for ADHD and scoliosis and XGBoost with Clustering-2 and Clustering-3 for scoliosis. However, mapping schema outperformed unsupervised feature reduction methods for most sub-phenotypes and algorithms.

Phecode-based feature sets Phecode-1 and Phecode-2 delivered the highest performances across the 3 sub-phenotypes. Models using these feature sets were the only ones to perform above average by nearly every metric for every sub-phenotype. Across sub-phenotypes, both Phecode-1 and Phecode-2 had the highest overall performance with XGBoost relative to the other algorithms. The AUROC for the Phecode-1 XGBoost model was 0.77 for OPG, 0.68 for ADHD, and 0.69 for scoliosis. The AUROC for the Phecode-2 XGBoost model was 0.78 for OPG, 0.68 for ADHD, and 0.71 for scoliosis.

Phecode-3 performed poorly across the 3 sub-phenotypes by nearly every metric. The best-performing Phecode-3 algorithm was logistic regression, with an AUROC of 0.68 for OPG, 0.64 for ADHD, and 0.60 for scoliosis.

Friedman's test to compare AUROCs of each feature set across models and conditions was statistically significant ( $P$ -value = 0.0034). Subsequent post hoc analysis using the Conover test identified significant differences between the following pairs: ICD10-3 and Phecode-3 ( $P$ -value = 0.031), Phecode-1 and Phecode-3 ( $P$ -value = 0.018), and Phecode-2 and Phecode-3 ( $P$ -value = 0.016).

### Feature importance overlap

For OPG using the XGBoost algorithm, the feature set producing the most shared features as ICD10-1 in the model was ICD10-2, which shared all the top 10 features. The model

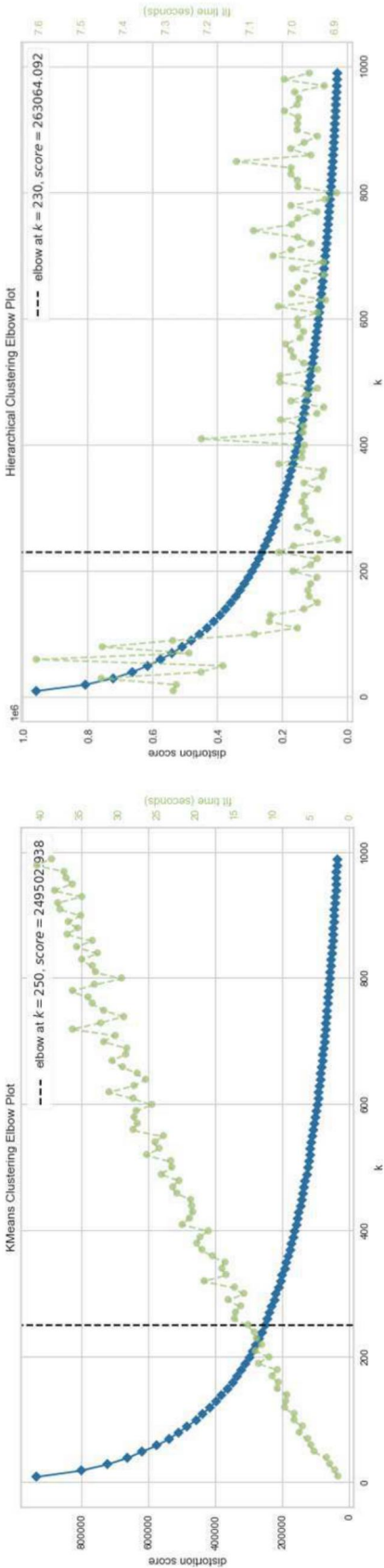
using the Clustering-2 feature set was the only model that did not share any important features with ICD10-1. The counts of shared features between models can be found in Table S2. As shown in Figure 4, the plot is dense, with several feature sets having at least half of their features as shared features, indicating that, for the prediction of OPG using the XGBoost algorithm, many of the feature sets produced multiple shared features with ICD10-1.

For ADHD using the XGBoost algorithm, feature sets ICD10-2, ICD10-3, and Phecode-2 had the most shared features, with ICD10-1 at 4 features. The models using Clustering-1, Clustering-2, and Clustering-3 feature sets did not share any important features with ICD10-1. The counts of shared features between models can be found in Table S3. As shown in Figure 5, the absence of any shared features for 3 of the features set and the relatively small number of shared features for those that do cause the plot to be relatively sparser than OPG as shown in Figure 4.

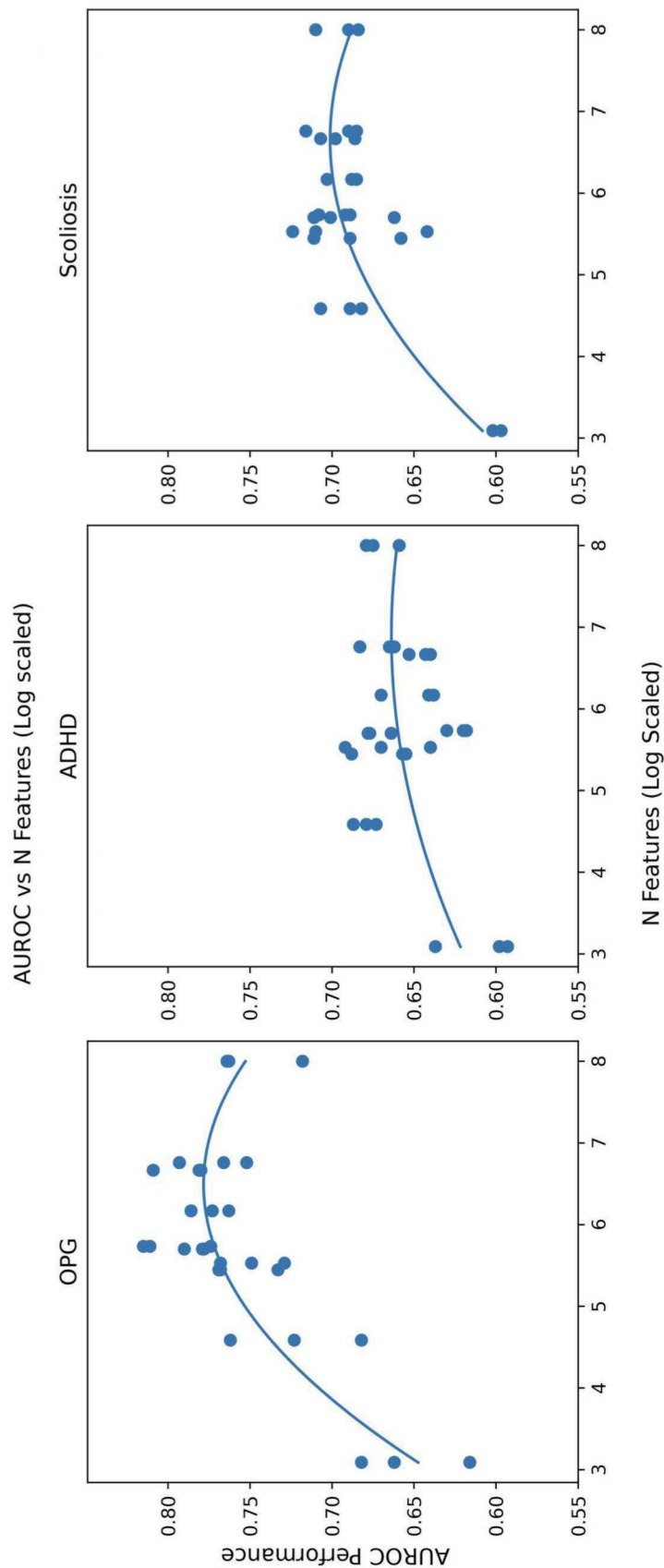
For scoliosis using the XGBoost algorithm, feature sets Phecode-1, Phecode-2, Phecode-3, and CCSR had the most shared features with ICD10-1 at 3 features. The models using Clustering-1, Clustering-2, and Clustering-3 did not share any important features with ICD10-1. The counts are included in Table S4. As shown in Figure 6, the plot is sparser than that observed for OPG or ADHD and has no feature set with over 2 shared features.

### Discussion

Early prediction of risk for developing clinical sub-phenotypes of NF1 would allow for better monitoring, early intervention, and improved outcomes and quality of life for affected children and their families. In one of our previous study, we used data from multiple clinical databases to successfully recapitulated several important and clinically relevant patterns in NF1 semiology specifically based on demographic and clinical characteristics.<sup>2</sup> We also developed applied ML techniques to predict sub-phenotypes in NF1 using data from multiple clinical datasets including a manually curate clinical database. However, maintaining longitudinal clinical databases is challenging due to the significant time, cost, and human effort required. They may also introduce inconsistencies in data or format over time, resulting in a lack of reproducibility of findings derived from these cohorts. On the other hand, EHR constitutes a rich source of longitudinal real-world clinical data accrued automatically during routine healthcare encounters, making it a powerful resource for predicting health outcomes.<sup>26-28</sup> High-dimensional health datasets like the EHR are often



**Figure 2.** Left elbow plot for K-means clustering. Right elbow plot for hierarchical clustering. The dashed black line represents the optimal number of clusters in both the plots.

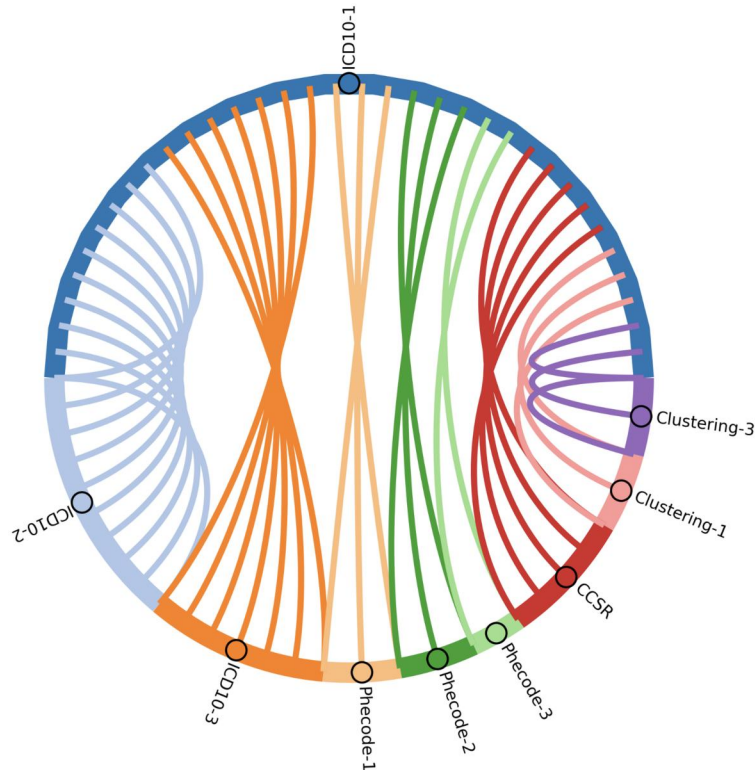


**Figure 3.** Model AUROC plotted against number of features (log-scaled) in the machine learning models for OPG (left), ADHD (middle), and scoliosis (right). The second degree polynomial line of best fit shows the AUROC performance as a function of log scaled number of features. The far right of the x-axis shows the feature sets without any feature reduction, and moving to the left along the x-axis shows the feature sets with more feature reduction. The peak in the best-fit line at the middle of the x-axis shows that some feature reduction results in a higher AUROC than the non-reduced feature sets. However, moving past the peak to feature sets with more feature reduction shows a drop in AUROC.

**Table 3.** Model performance.

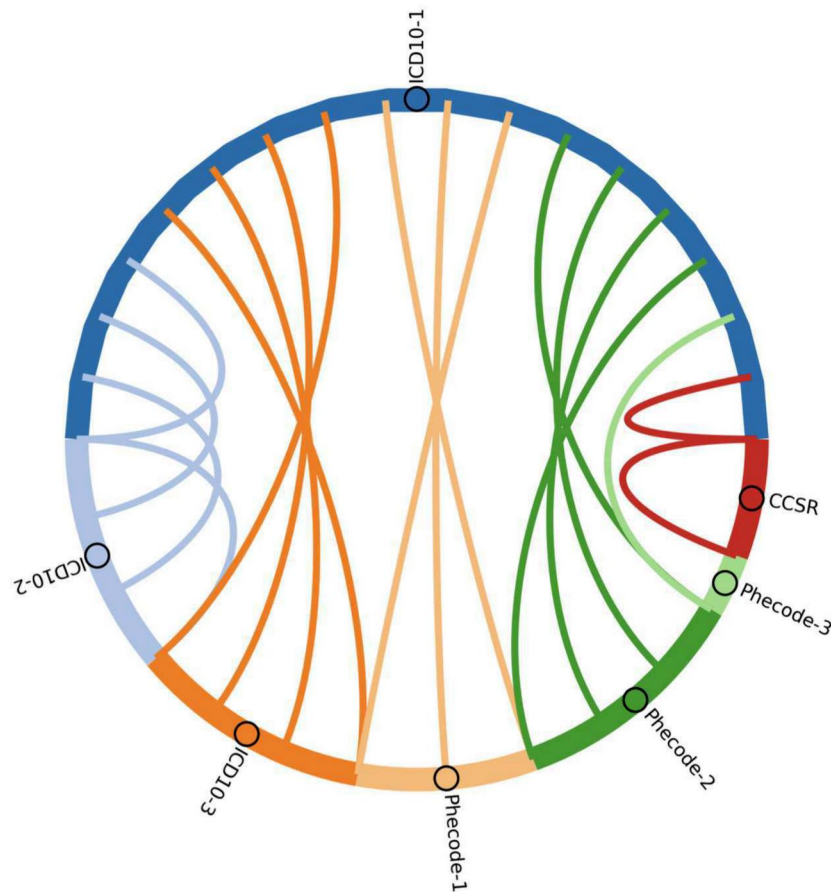
Model	Feature set (N)	OPG		ADHD		Scoliosis	
		AUROC	F1	AUROC	F1	AUROC	F1
Logistic regression	ICD10-1 (2982)	<b>0.76</b>	<b>0.42</b>	<b>0.68</b>	<b>0.48</b>	<b>0.68</b>	<b>0.41</b>
	ICD10-2 (477)	0.77	0.39	0.64	0.47	<b>0.69</b>	<b>0.40</b>
	ICD10-3 (786)	<b>0.81</b>	<b>0.49</b>	0.64	0.47	<b>0.69</b>	<b>0.41</b>
	Phencode-1 (861)	<b>0.79</b>	<b>0.50</b>	0.67	0.47	0.69	0.39
	Phencode-2 (299)	0.78	0.45	<b>0.68</b>	<b>0.50</b>	0.66	0.39
	Phencode-3 (22)	0.68	0.34	0.64	0.49	0.60	0.34
	CCSR (309)	<b>0.82</b>	<b>0.48</b>	0.63	0.48	0.69	0.41
	Clustering-1 (252)	0.77	0.42	0.64	0.46	0.64	0.35
	Clustering-2 (232)	0.77	0.42	0.66	0.48	0.66	0.37
	Clustering-3 (98)	0.76	0.43	<b>0.69</b>	<b>0.49</b>	<b>0.71</b>	<b>0.42</b>
Random forest	ICD10-1 (2982)	0.72	0.39	0.68	0.50	<b>0.71</b>	<b>0.42</b>
	ICD10-2 (477)	<b>0.79</b>	<b>0.49</b>	0.67	0.51	0.70	0.43
	ICD10-3 (786)	<b>0.78</b>	<b>0.47</b>	0.64	0.46	0.70	0.40
	Phencode-1 (861)	0.75	0.43	0.66	0.48	<b>0.72</b>	<b>0.36</b>
	Phencode-2 (299)	<b>0.79</b>	<b>0.46</b>	0.66	0.49	0.70	0.41
	Phencode-3 (22)	0.62	0.16	0.59	0.34	0.60	0.24
	CCSR (309)	0.77	0.44	0.62	0.46	0.69	0.37
	Clustering-1 (252)	0.73	0.44	<b>0.69</b>	<b>0.46</b>	<b>0.72</b>	<b>0.35</b>
	Clustering-2 (232)	0.73	0.45	<b>0.69</b>	<b>0.44</b>	0.69	0.29
	Clustering-3 (98)	0.68	0.32	<b>0.68</b>	<b>0.36</b>	0.68	0.20
XGBoost	ICD10-1 (2982)	0.76	0.41	0.66	0.51	0.69	0.39
	ICD10-2 (477)	0.76	0.45	0.64	0.51	0.69	0.40
	ICD10-3 (786)	<b>0.78</b>	0.48	0.65	0.50	0.71	0.38
	Phencode-1 (861)	0.77	0.46	<b>0.68</b>	<b>0.51</b>	0.69	0.40
	Phencode-2 (299)	<b>0.78</b>	<b>0.47</b>	<b>0.68</b>	<b>0.52</b>	<b>0.71</b>	<b>0.41</b>
	Phencode-3 (22)	0.66	0.36	0.60	0.49	0.60	0.32
	CCSR (309)	<b>0.81</b>	<b>0.54</b>	0.62	0.48	0.71	0.40
	Clustering-1 (252)	0.75	0.41	0.67	0.48	<b>0.71</b>	<b>0.40</b>
	Clustering-2 (232)	0.77	0.47	0.66	0.48	<b>0.71</b>	<b>0.40</b>
	Clustering-3 (98)	0.72	0.42	<b>0.67</b>	<b>0.49</b>	0.69	0.39

BOLD: Top 3 results per model per condition are bolded.  
Abbreviations: OPG = Optic Pathway Glioma; ADHD = Attention-Deficit/Hyperactivity Disorder; ICD = International Classification of Diseases; ICD10-CM = International Classification of Diseases, Tenth Revision, Clinical Modification; CCSR = Clinical Classifications Software Refined; PCA = Principal Component Analysis.



**Figure 4.** OPG feature overlap showing the shared important features between ICD10-CM Base and the rest of the feature set important features. Each node along the outer circle represents a feature set, and each edge connecting ICD10-CM Base to another feature set represents one shared feature. For example, if ICD10-CM Base and CCSR have 5 shared features, there would be 5 edges connecting their nodes. The denser the plot, the greater similarity of the most important features of the feature space for that outcome variable. Inversely, the sparser the plot, the less similar the most important features of the feature space are. Feature sets producing no shared features with ICD10-CM Base are not shown in the plot.





**Figure 5.** ADHD feature overlap showing the shared important features between ICD10-CM Base and the rest of the feature set important features. Each node along the outer circle represents a feature set, and each edge connecting ICD10-CM Base to another feature set represents 1 shared feature. For example, if ICD10-CM Base and CCSR have 5 shared features, there would be 5 edges connecting their nodes. The denser the plot, the greater similarity of the most important features of the feature space for that outcome variable. Inversely, the sparser the plot, the less similar the most important features of the feature space are. Feature sets producing no shared features with ICD10-CM Base are not shown in the plot.

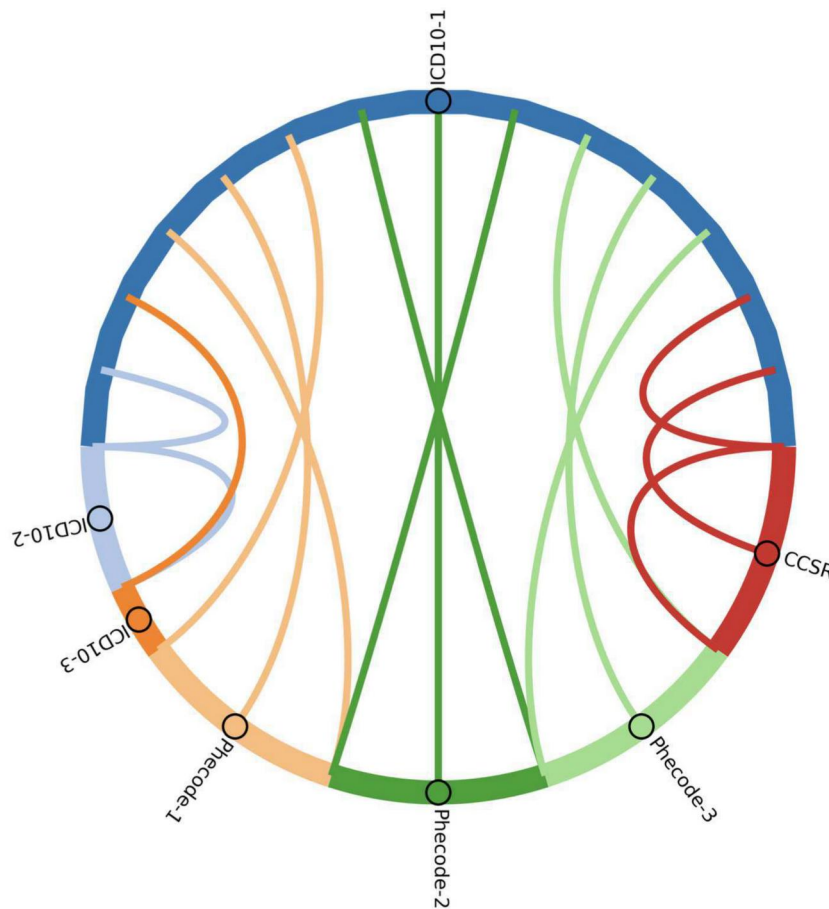
challenging to use in AI-based models due to large raw input and small sample size, which is particularly true in the context of rare-disease research. Performance and generalizability of predictive models can be improved by selecting high-level and important features.

Using the prediction of NF1 sub-phenotypes as a case study, we examined how different dimensionality reduction strategies can affect the performance of a predictive model. Our experiments revealed that our models were most successful at predicting OPG, followed by scoliosis and ADHD. This may be because, from a clinical standpoint, there is less ambiguity regarding the diagnosis of OPG compared to ADHD, as suggested by previously published studies.<sup>2</sup> Across the predicted sub-phenotypes, XGBoost was the best-performing algorithm overall, while the best-performing feature sets were generally derived from domain knowledge-informed mapping schema. However, Phecode-3 (Phecode Exclusive), which represented the highest-level Phecode categorizations and resulted in the fewest number of very-high-level features, was the worst-performing feature set across all algorithms and predicted outcomes, indicating excessive information loss. Statistically significant comparisons between Phecode-3 and various feature sets illustrate that the performance of ML models declines with excessive information loss when using high-level features.

Looking specifically at the best combination of algorithm and aggregated/reduced feature set varied for each sub-

phenotype: the best model for predicting OPG was the combination of the XGBoost algorithm with the CCSR feature set, the best models for predicting scoliosis were the random forest and XGBoost algorithms with the ICD10-1 and Phecode-2 feature sets respectively, and the best model for predicting ADHD was the XGBoost algorithm with the Phecode-2 feature set. In conclusion, overall, XGBoost was the best-performing model, and feature aggregation methods based on Phecodes-2 performed well across all conditions. In addition, the prediction of OPG, which also had the best-performing models, had the most important features shared across the feature sets. This may suggest that comorbidities are a much stronger predictor for OPG than for ADHD and scoliosis. The overlapping features may also represent potential factors that could be used as future clinical risk stratification markers.

These findings demonstrate that in model development, besides optimizing algorithm parameters, it may also be necessary to experiment with several different approaches to feature aggregation and reduction to arrive at the optimal model. While this may require more computational resources, the feature aggregation and reduction methods described here can be automated, require little manual effort and no domain expertise (beyond the initial creation of the mapping schema), and thus represent a highly feasible model optimization step. Because the scope of our study was limited to comparing the effect of feature reduction and aggregation of



**Figure 6.** Scoliosis feature overlap showing the shared important features between ICD10-CM Base and the rest of the feature set important features. Each node along the outer circle represents a feature set, and each edge connecting ICD10-CM Base to another feature set represents 1 shared feature. For example, if ICD10-CM Base and CCSR have 5 shared features, there would be 5 edges connecting their nodes. The denser the plot, the greater similarity of the most important features of the feature space for that outcome variable. Inversely, the sparser the plot, the less similar the most important features of the feature space are. Feature sets producing no shared features with ICD10-CM Base are not shown in the plot.

diagnosis codes, for which well-validated mapping schema already exists, the approach of using domain knowledge-based mapping for dimensionality reduction may not be generalizable depending on the types of features being used in a model. In the future, we plan to expand our feature aggregation and reduction to include additional EHR data elements, such as laboratory results and medications.

## Conclusion

Dimensionality reduction through feature aggregation can improve the performance of ML models, especially in high-dimensional datasets with small sample sizes, often seen in healthcare. However, it may result in information loss and data oversimplification without careful optimization, potentially negatively impacting model performance. The choice of technique for dimensionality reduction can affect model performance and varies by the ML algorithm used and the outcome being predicted. During model development, it is important to experiment with several feature selection methods to optimize model performance. Future directions for this work involve temporal modeling at the encounter level rather than the patient level and expanding the data to include labs, vitals, procedures, etc., with their corresponding concept mappings. Also, applying this study to a condition besides NF1 would provide valuable insights into

performance patterns of algorithms and dimensionality reduction techniques that speak to their generalizability.

## Author contributions

Aditi Gupta (Conceptualization, Investigation, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing—original draft, Writing—review and editing), Ethan Hillis (Conceptualization, Investigation, Methodology, Software, Writing—original draft, Validation, Writing—review and editing, Visualization), Inez Y. Oh (Conceptualization, Data curation, Methodology, Project administration, Validation, Writing—original draft, Writing—review and editing), Stephanie M. Morris (Methodology, Validation, Writing—review and editing), Zach Abrams (Methodology, Validation, Writing—review and editing), Randi E. Foraker (Methodology, Validation, Writing—review and editing), David H. Gutmann (Methodology, Validation, Writing—review and editing), and Philip R.O. Payne (Conceptualization, Project administration, Funding acquisition, Supervision, Writing—review and editing)

## Supplementary material

[Supplementary material](#) is available at *JAMIA Open* online.

## Funding

This work was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number R01NS131112.

## Conflicts of interest

The authors declare no competing interests.

## Data availability

The patient level data underlying this article cannot be shared publicly in order to protect patient privacy. The data will be shared on reasonable request to the corresponding author.

## References

- Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *NPJ Digit Med*. 2021;4:153.
- Morris SM, Gupta A, Kim S, et al. Predictive modeling for clinical features associated with neurofibromatosis type 1. *Neurol Clin Pract*. 2021;11:497-505.
- Mukherjee P, Humbert-Droz M, Chen JH, et al. SCOPE: predicting future diagnoses in office visits using electronic health records. *Sci Rep*. 2023;13:11005.
- Deschepper M, Waegeman W, Vogelaers D, et al. Using structured pathology data to predict hospital-wide mortality at admission. *PLoS One*. 2020;15:e0235117.
- Brämer GR. International statistical classification of diseases and related health problems. Tenth revision. *World Health Stat Q*. 1988;41:32-36.
- ICD-10-CM official guidelines for coding and reporting FY 2023—Updated April 1, 2023 (October 1, 2022 - September 30, 2023). 2023. Accessed January 10, 2025. <https://stacks.cdc.gov/view/cdc/126426>.
- Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Med Inform*. 2019;7:e14325.
- Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses. Accessed January 10, 2025. <https://hcupus.ahrq.gov/toolssoftware/ccsr/dxcsr.jsp>
- Rasmy L, Tiriyaki F, Zhou Y, et al. Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. *J Am Med Inform Assoc*. 2020;27:1593-1599.
- Liu N, Chee ML, Koh ZX, et al. Utilizing machine learning dimensionality reduction for risk stratification of chest pain patients in the emergency department. *BMC Med Res Methodol*. 2021;21:74.
- Gutmann DH, Ferner RE, Listernick RH, et al. Neurofibromatosis type 1. *Nat Rev Dis Primers*. 2017;3:17004.
- Ferner RE, Gutmann, DH. Chapter 53 - Neurofibromatosis type 1 (NF1): diagnosis and management. In: Said G, Krarup C, eds. *Handbook of Clinical Neurology*. Elsevier; 2013:939-955.
- Jabaudon M, Blondonnet R, Audard J, et al. Recent directions in personalised acute respiratory distress syndrome medicine. *Anaesth Crit Care Pain Med*. 2018;37:251-258.
- King A, Listernick R, Charrow J, et al. Optic pathway gliomas in neurofibromatosis type 1: the effect of presenting symptoms on outcome. *Am J Med Genet A*. 2003;122a:95-99.
- Cohen R, Halevy A, Aharon S, et al. Attention deficit hyperactivity disorder in neurofibromatosis type 1: evaluation with a continuous performance test. *J Clin Neurol*. 2018;14:153-157.
- Konieczny MR, Senyurt H, Krauspe R. Epidemiology of adolescent idiopathic scoliosis. *J Child Orthop*. 2013;7:3-9.
- Jin X, Han, J. K-means clustering. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning*. Springer US; 2010:563-564.
- Jolliffe I. Principal component analysis. In: Lovric M, ed. *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg; 2011:1094-1096.
- Bengfort B, Bilbro R, Danielsen R, et al. Yellowbrick. Accessed January 10, 2025. <https://www.scikityb.org/en/latest/index.html>
- McCullagh P, Nelder J. *Generalized Linear Models*/P. McCullagh, J.A. Nelder. SERBIULA (sistema Librum 2.0). 1986, 28.
- Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
- Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016:785-794.
- Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. *Sci Rep*. 2024;14:6086.
- Conover WJ. *Practical Nonparametric Statistics*, 3rd ed. Wiley Series in Probability and Statistics. Wiley; 1999.
- Hansen SH. holoviz/holoviews: Version 1.16.2 (v1.16.2). Zenodo. 2023. Accessed January 10, 2025. <https://holoviews.org/about.html>.
- Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep*. 2017;7:46226.
- Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011;3:79re1.
- Wei W-Q, Teixeira PL, Mo H, et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc*. 2016;23:e20-7-e27.