Evaluation of standard and semantically-augmented distance metrics for neurology patients

Daniel B Hier (dhier@uic.edu)
University of Illinois at Chicago College of Medicine https://orcid.org/0000-0002-6179-0793
Jonathan Kopel
Texas Tech University Health Sciences Center School of Medicine
Steven U Brint
University of Illinois at Chicago College of Medicine
Donald C Wunsch II
Missouri University of Science and Technology
Gayla R Olbricht
Missouri University of Science and Technology
Sima Azizi
Missouri University of Science and Technology
Blaine Allen
Missouri University of Science and Technology

Research article

Keywords: Patient distances, Semantic augmentation, Ontologies, Machine learning, Patient clustering, Patient classification, Distance metrics, neurology

Posted Date: August 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-20018/v4

Version of Record: A version of this preprint was published on August 26th, 2020. See the published version at https://doi.org/10.1186/s12911-020-01217-8.

Evaluation of standard and semantically-augmented distance metrics for neurology patients

Daniel B. Hier[a], Jonathan Kopel[b], Steven U. Brint[a], Donald Wunsch II[c], Gayla R. Olbricht[d], Sima Azizi[c], Blaine Allen[c]

[a]Department of Neurology and Rehabilitation, University of Illinois at Chicago, Chicago IL

[b]Department of Internal Medicine, Texas Tech University Health Sciences Center, Lubbock TX

[c]Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla MO 65401

[d]Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla MO 65401

Corresponding Author:

Daniel B. Hier MD

Department of Neurology and Rehabilitation

University of Illinois at Chicago

Chicago Il 60612

Email: dhier@uic.edu

Telephone: (312) 622-2776

Daniel B. Hier
ORCHID ID: https://orcid.org/0000-0001-9603-0797

Donald C. Wunsch II
ORCID ID: https://orcid.org/0000-0002-9726-9051

Jonathan Kopel
Orchid ID: https://orcid.org/0000-0001-5934-2695

Blaine Allen
Orchid ID: https://orcid.org/0000-0001-9603-0797

Gayla R. Olbricht
Orchid ID: https://orcid.org/0000-0002-1213-2241

Abstract

Background:

Patient distances can be calculated based on signs and symptoms derived from an ontological hierarchy. There is controversy as to whether patient distance metrics that consider the semantic similarity between concepts can outperform standard patient distance metrics that are agnostic to concept similarity. The choice of distance metric can dominate the performance of classification or clustering algorithms. Our objective was to determine if semantically augmented distance metrics would outperform standard metrics on machine learning tasks.

Methods:

We converted the neurological findings from 382 published neurology cases into sets of concepts with corresponding machine-readable codes. We calculated patient distances by four different metrics (cosine distance, a semantically augmented cosine distance, Jaccard distance, and a semantically augmented bipartite distance). Semantic augmentation for two of the metrics depended on concept similarities from a hierarchical neuro-ontology. For machine learning algorithms, we used the patient diagnosis as the ground truth *label* and patient findings as machine learning *features*. We assessed classification accuracy for four classifiers and cluster quality for two clustering algorithms for each of the distance metrics.

Results:

Inter-patient distances were smaller when the distance metric was semantically augmented. Classification accuracy and cluster quality were not significantly different by distance metric.

Conclusion:

Although semantic augmentation reduced inter-patient distances, we did not find improved classification accuracy or improved cluster quality with semantically augmented patient distance metrics when applied to a dataset of neurology patients. Further work is needed to assess the utility of semantically augmented patient distances.

Keywords:
Patient distances
Semantic augmentation
Ontologies
Machine learning
Patient clustering
Patient classification
Distance metrics
neurology

Background and Related Work

Patients present with signs (what the physician finds on examination) and symptoms (patient complaints). We group *signs and symptoms* under the more general term *findings* [1]. Distance metrics play an important role in advancing precision medicine, machine learning, and patient phenotyping [2-12]. Patient distances can be calculated based on findings that have been converted to machine codes based on concepts from a hierarchical ontology.

$$signs + symptoms = findings \rightarrow concepts \rightarrow machine\ codes$$

In this study, we examine whether the semantic augmentation of distance metrics with concept similarities improves the classification and clustering of neurology patients.

Distance metrics.

A variety of similarity and distance metrics are available. These have been used to calculate distances between patients [13-16], documents [17-19], and phenotypes [4, 5, 9, 10, 12]. If similarity and distance metrics are normalized to a scale of 0.0 to 1.0, the distance between A and B is the complement of the similarity.

$$\text{distance}(A, B) = 1 - \text{similarity}(A, B). \quad (1)$$

The distance between two patients is different than the distance between two medical concepts. Patients are complex and can be represented as a collection of many concepts. Inter-patient distances are many-to-many comparisons; inter-concept distances are one-to-one comparisons. Metrics that work for concept distances are generally different from metrics to calculate distances between patients. Melton et al. [16] comment that "semantic distance measures the relative closeness between two concepts….Inter-patient distance compares the relative closeness between two cases (sets of patient data)."

The implementation of distance metrics for neurological patients based on findings is challenging. First, neurological findings are recorded as unstructured free text. Second,

examiners use a variety of equivalent terms to represent the same meaning: hyperreflexia is equivalent to increased reflexes; Babinski sign is equivalent to extensor plantar response; and so on. Third, the number of findings may vary from patient to patient. Fourth, converting unstructured text into machine-readable codes is difficult [20-21].

The SNOMED CT ontology and the UMLS Metathesaurus allow the consolidation of multiple synonymous terms under the same concept [22-23]. Both terminologies assign unique machine-readable codes to a concept. We have identified 1204 core concepts from the UMLS Metathesaurus as a neuro-ontology for capturing findings of the neurological examination [24]. This curated neuro-ontology has three characteristics that make it well-suited for patient distance calculations: 1) it is monohierarchic, 2) the neurologic similarity of concepts has organized its hierarchy, and 3) it contains neurologic concepts absent from SNOMED CT [24].

When findings are converted to concepts and represented as machine-readable codes, patients can be instantiated mathematically as a set (an unordered collection of findings) or as a vector (ordered array of elements of fixed length). If a patient is represented as a set, each finding is added to the set as a unique element. The cardinality of the set (number of set elements) is equal to the number of findings. If a patient is represented as a vector, each finding is represented as an element of the vector. The number of elements is equal to the number of potential findings. A variety of distance metrics can be used with vectors, including Manhattan, Euclidean, cosine, Pearson correlation, Hamming, Minkowski, and others [25]. Commonly used distance metrics in patient similarity studies are Jaccard, Mahalanobis, Euclidean, and cosine 15, 26]. Haase et al. [27] have suggested a bipartite matching algorithm for set similarity (equation 2) where |A| is the number of elements in $\mathrm{set\ A}$ and $\mathrm{sim(a, b)}$ is the similarity between a concept *a* from $\mathrm{set}\ A$ and *b* is a concept from $\mathrm{set\ B}$.

$$Sim\ (A, B) = \frac{1}{|A|} * \sum_{a \in A} max_{b \in B}\ (sim(a, b)). \quad (2)$$

Bipartite similarity metrics resembling equation 2 have been used to calculate patient distances [16].

Hierarchical ontologies such as SNOMED CT and the UMLS Metathesaurus allow the calculation of distances between concepts [28-36]. Concept distances derived from hierarchal ontologies show modest correlations with the distance judgments of human experts [35, 37, 38]. The distance metrics for both sets and vectors can be augmented by considering the similarity between concepts [13, 14, 19]. Melton et al. [16] compared computed patient distances with an expert opinion on patient distance based on chart review. They did not find that semantic augmentation of the distance metric enhanced correlation with expert opinion and that correlation between experts and computed patient distances was low regardless of semantic augmentation. Mabotuwana et al. [19] examined document similarity using a cosine distance metric after converting document concepts to a binarized vector. In a classification task that involved determining whether a radiological report was a head CT scan or an abdomen CT scan, they found the accuracy of a k-nearest neighbor classifier increased from 86.7% to 93.1% with semantic augmentation of the document vector based on the SNOMED CT concept hierarchy. Mabotuwana et al. found that semantic augmentation of inter-document distances increased the separation between the centroid of the head CT scan reports and the centroid of the abdomen CT reports. Jia et al. [14] examined the ability of patient distances generated by ICD-10 diagnoses to predict hospital length of stay. Although they explored a variety of distance metrics, including cosine, Jaccard, and bipartite matching, they came to no definite conclusion as to whether semantic augmentation (based on a concept hierarchy) improved classification accuracy. In the Human Phenotype Ontology (HPO), Kohler et al. [12] have implemented a semantically augmented distance metric to assist in matching unknown patients to archetypical patients in the Online Mendelian Inheritance in Man (OMIM) database. Girardi et al. [13] calculated distances between patients with diseases of the gall bladder, thyroid, or appendix and hernias based on ICD-10 diagnosis codes. They found that a semantically augmented patient distance metric outperformed a Jaccard distance on a clustering

task and that a semantically augmented patient distance increased the distance between within-diagnosis centroids and between diagnosis centroids.

Machine Learning.

Machine learning is increasingly used in the analysis of patient data. Machine learning is divided into supervised and unsupervised learning [39]. The prototypical tasks for supervised learning are classification and regression [40]. Although there are many machine learning classifiers, some commonly used classifiers include naïve Bayes, logistic regression, k-nearest neighbor, and random forest [40]. Naïve Bayes utilizes probabilities derived from predictor variables to select class membership. Logistic regression is a statistical method that fits parameters to a logistic equation to predict class membership. k-nearest neighbor classifiers utilize distances between cases to predict class membership. Random forest classifiers use an ensemble of decision trees to predict class membership. The most common use of unsupervised learning algorithms is for the clustering of cases into homogeneous groups. Although many clustering algorithms are available, two of the most commonly used clustering algorithms are k-means clustering and agglomerative clustering [41]. Both of these algorithms utilize inter-case distances to form homogeneous clusters of cases. Indices of machine learning classification quality include precision, recall, F1, and accuracy [42]. Indices of machine learning clustering quality include homogeneity, completeness, Rand index, V-score, silhouette score [43-45]. Distance metrics are frequently used to generate patient distance matrices that drive the clustering or classification of patients. Since the performance of machine learning clustering and classification algorithms can be assessed objectively, we have hypothesized that the semantic augmentation of distance metrics with inter-concept distances would improve the performance of these algorithms.

To test this hypothesis, we created four test groups of patients abstracted from textbooks. We investigated four classifiers (naïve Bayes, logistic regression, random forests, and k-nearest neighbor) and two clustering algorithms (agglomerative and k-means) across four distance metrics. We tested whether semantic augmentation of the distance metrics improved clustering or classification quality.

Methods

Case abstraction.

We created a dataset of 382 neurological patients selected from a convenience sample [46] of 1028 published teaching cases [47-58]. We abstracted 2616 findings from the case studies (mean 6.7 ± 3.4 findings per patient). Findings were transcribed verbatim from source materials. An abstractor manually selected one of the 1204 available terms in the neuro-ontology that best represented the finding and added the UMLS CUI code [24]. Table 1 illustrates the case abstraction method for a patient with Parkinson disease.

Distance metrics.

We implemented four inter-patient distance metrics in Python [59]. The *Jaccard distance* is the complement of the Jaccard similarity [60]. If A and B are the sets of findings from patient A and patient B, the $\text{Jaccard}_{\text{dist}}$ (A, B) is shown by equation (3), and J_{sim} is the Jaccard similarity.

$$\text{Jaccard}_{dist}(\text{A}, \text{B}) = 1 - \text{J}_{\text{sim}}(\text{A}, \text{B}) = 1 - \frac{\text{A} \cap \text{B}}{\text{A} \cup \text{B}}. \quad \textbf{(3)}$$

The *augmented bipartite distance* is based on the metric of Melton et al. [16] after augmenting it with the inter-concept distance proposed by Wu and Palmer [29]. If patients A and B are represented as a set of findings such that a ϵ A and b ϵ B, the augmented bipartite distance is shown by equation (4) and is supported by equations (5), (6), and (7).

$$\text{agumented bipartite distance } (\text{A}, \text{B}) = \frac{\text{D } (\text{A}, \text{B}) + \text{D } (\text{B}, \text{A})}{2}. \quad \textbf{(4)}$$

$$D(A, B) = \frac{1}{|A|} * \sum_{a \epsilon A} min_{b \epsilon B} \text{ dist } (\text{a}, \text{b}). \quad \textbf{(5)}$$

$$D(B, A) = \frac{1}{|B|} * \sum_{b \epsilon A} min_{a \epsilon B} \, dist \, (a, b). \quad \textbf{(6)}$$

$$dist \, (a, b) = \; 1 - \frac{2 * depth(LCS)}{depth \, (a) + depth(b)}. \quad \textbf{(7)}$$

For equation (7), we used the hierarchical structure of the neuro-ontology and the method of Wu and Palmer [29] to calculate the *dist (a, b)* as the semantic distance between concept *a* and concept *b*. *LCS* is the lowest common subsumer in the hierarchical ontology for concepts a and b; *depth(a)* is the number of levels from the root concept to concept a; *depth (b)* is the number of levels from the root concept to concept *b*, and *depth(LCS)* is the number of levels from the root concept to the *LCS*. Based on equation (7), the dist (a, b) for each inter-concept distance was stored as a *nxn* lookup table where the number of possible concepts was n = 1204. Values from this lookup table were used in equations (5) and (6) to iteratively find the minimum inter-concept distance for each concept from patient A compared to the concepts in patient B. *Cosine distances* between patients (1 – cosine similarity) were calculated by standard methods (equation 8). If patient A and patient B are represented as vectors of findings from a_1 to a_n and from b_1 to b_n, the vector is binarized, so that a_i or b_i is 1 if the finding is present and 0 if the finding is absent. Patient vectors were represented as a one-dimensional array of length n= 1204, where n is the potential number of findings.

$$cosine\ distance(A,B) = 1 - \frac{\sum(a_i\ * b_i\ )}{(\sqrt{\sum \mathrm{a_i^2}}) * (\sqrt{\sum \mathrm{b_i^2}})}. \quad \textbf{(8)}$$

We calculated an *augmented cosine distance* between patients according to the method of Mabotuwana et al. [19] Patients were represented as one-dimensional arrays as in the cosine distance above. We used the hierarchical structure of the neuro-ontology [24] to find an ordered list of ancestors for each concept. For each of the 1204 concepts in the neuro-ontology, we created a semantically augmented vector. The formula for augmentation was 1/(1+n) where n = 0 for the index concept, n=1 for the parent concepts, n=2 for the grandparent concepts, etc. Descendent concepts (children) in the neuro-ontology were not augmented. Ancestor hierarchy was determined by the neuro-ontology, which is mono-hierarchical [24]. Augmentation vectors were stored in an n*xn* lookup table (n=1204). Semantically augmented patient vectors were created for each patient by traversing a list of concepts for each patient and adding the augmented

concept vector to the patient vector to obtain a summary patient vector. After semantic augmentation of the vectors, inter-patient distances were calculated by equation 8.

For all metrics, distances were positive, symmetric, and normalized between 0.0 and 1.0. Distances for each distance metric were stored in a square *nxn* matrix (n = 382 patients) before input to classification or clustering algorithms.

Test Groups

We divided the dataset of 382 patients into four test groups by diagnosis (Table 2). Each test group consisted of patients with eight related diagnoses. Each diagnosis occurred at least four times (mean 11.9 ± 5.9) in the test group. Test groups were composed of competing diagnoses for a common presenting neurological complaint (*a patient with weakness, a patient with abnormal movements, a patient with altered mental status, and a patient with cranial neuropathy*). Diagnoses were selected to emulate the differential diagnosis a neurologist might consider when evaluating a patient complaint.

Classification and Clustering

For the classification tasks, we assessed the ability to assign correctly *diagnoses* based on *findings*. The ground truth labels were the diagnoses from the abstracted patient histories, and the features were the abstracted findings. Naïve Bayes, logistic regression, random forest, and k-nearest neighbor classifiers were compared. We used the Orange 3.25 default hyperparameters for naïve Bayes. For logistic regression, we set regularization = L2, and for random forest, we set the number of trees = 10. For the k-nearest neighbor classifier, we used uniform distance weighting and k=5 after the empirical evaluation of all k values between 2 and 15. We used classification accuracy and a balanced F1 score to assess classification performance based on 10-fold cross-validation [42]. In a separate analysis, we found mean F1 scores and mean accuracy scores did not differ statistically (df =1, p>.05) between the 10-fold cross-validation method and the random sampling validation method.

For both the agglomerative clustering algorithm (Ward linkage) [61] and the k-means clustering algorithm, we chose a hyperparameter of *number of clusters = 8* based on the known number of diagnoses in the test groups (Table 2). We used the silhouette score, homogeneity score, completeness score, V-score, adjusted Rand index, and mutual information index to assess cluster quality [42-45, 59].

Statistical methods.

We used SPSS 26 (IBM Corporation) for analysis of variance, line plots, and box plots. We used Orange 3.25.0 for the k-nearest neighbor, logistic regression, naïve Bayes, and random forest classifications. We used scikit-learn 0.23.1 for agglomerative clustering and k-means clustering [59]. All performance measures for clustering and classification were normalized to a 0 to 100 scale.

Results

We examined inter-patient distances for 382 patients divided into 4 test groups of eight diagnoses (Table 2). Inter-patient means differed by distance metric (Figure 1, one-way ANOVA, df =3, F= 5820, $p < .001$). Post hoc means testing (Bonferroni $p < .05$) showed all means differed ($p < .05$) with the augmented bipartite distance metric having the lowest inter-patient mean distance and the Jaccard distance metric having the highest mean inter-patent distance.

The mean within-diagnosis patient distance was less than mean between-diagnosis patient distance for all the four-distance metrics (Figure 2, two-way ANOVA, means differ by group, df =1, F=3050, $p < .001$ and means differ by distance metric, df = 3, F=2936, $p < .001$). All pairwise mean comparisons by the group and by distance metric were significant (post hoc Bonferroni test, $p < .05$).

We found a significant difference in mean patient distances by diagnosis (Figure 3, two-way ANOVA, means differ by diagnosis, df =31, F=107, $p < .001$, and means differ by distance metric, df = 3, F=1351, $p < .001$). Post hoc Bonferroni testing showed that 60% of the pairwise patient distance means differed by diagnosis ($P < .05$). For the 32

diagnoses shown in Figure 3, trigeminal neuralgia has the lowest mean within-diagnosis patient distance (less than all other 31 diagnoses, pairwise comparisons, $p < .05$) and multiple sclerosis had the highest within-diagnosis mean patient distance (greater than all other diagnoses, pairwise comparisons, $p < .05$).

We performed 64 classification analyses (4 distance metrics x 4 test groups x 4 classifiers). The four test groups were *altered mental status*, *abnormal movement, cranial neuropathy, and weakness* (Table 2). The four distance metrics were cosine, augmented cosine, augmented bipartite, and Jaccard (see Methods). The four classifiers were naïve Bayes, logistic regression, random forest, and k-nearest neighbor (k=5). Classes were unbalanced in the test groups (Table 2). Each classification task involved selecting the correct diagnosis from one of eight competing diagnoses for each of the patients in the test group. The performance was measured by classification accuracy and F1. Classification performance varied by classifier for both classification accuracy (two-way ANOVA, main effect, $df=3$, $F =7.8$, $p < .001$) and F1 (two-way ANOVA, main effect, $dF=3$, $F=10.1$, $p < .001$). Bonferroni post hoc testing showed that the naïve Bayes classifier underperformed the logistic regression and k-nearest neighbor classifiers on both performance measures ($p < .05$).

Classification performance of the distance metrics was comparable regardless of classifier (Figures 4-5, two-way ANOVA, $df =3$, $p > .05$) or diagnosis group (two-way ANOVA, Figures 6-7, $df=3$, $p > .05$). Classifier performance was comparable when performance was measured by classification accuracy (Figures 4) or by F1 (Figure 5). Performance differed by diagnosis group (Figures 6 and 7) for both classification accuracy (two-way ANOVA, $df= 3$, $F=10.2$, $p < .001$) and the F1 score (two-way ANOVA, $df=3$, $F=7.4$, $P < .001$). Post hoc Bonferroni testing showed the classification accuracy score, and the F1 score was higher for the cranial nerve group than the other three diagnosis groups ($p < .05$).

We performed 32 clustering analyses (4 distance metrics x 4 test groups x 2 clustering algorithms). The two clustering algorithms were agglomerative clustering with Ward

linkage and k-means clustering. Distances were inputted as pre-computed *nxn* matrices. For both clustering algorithms, the number of clusters was set at eight based on the known number of different diagnoses in each diagnosis group. Cluster quality was assessed by silhouette score, adjusted Rand Index (ARI), adjusted mutual information (AMI), completeness, homogeneity, and V-measure. Cluster quality did not differ by cluster algorithm (agglomerative versus k-means) on any of the cluster quality measures (Figure 8, two-way ANOVA, df =1, $p>.05$).

For both k-means clustering and agglomerative clustering, the distance metric did not significantly affect cluster quality (Figures 9 and 10, two-way ANOVA, df=3, $p >.05$). Cluster quality was better for the cranial nerve group (Figure 11) than the other three groups, the movement group was better than the weakness group (Bonferroni post hoc test, $p <.05$; Groups differ two-way ANOVA, df= 3, $F=20.3$, $p <.001$). The higher quality of the cranial nerve clustering with greater within-cluster homogeneity than the weakness group clustering is illustrated in the stacked bar charts Figures 12 and 13.

Discussion

We examined four distance metrics for calculation of the distances between neurology patients based on findings: Jaccard distance, cosine distance, augmented cosine distance and augmented bipartite distance. To calculate the Jaccard and augmented bipartite distances, we represented patients as unordered lists of elements of variable length (sets). To calculate the cosine and augmented cosine distances, we represented patients as ordered arrays of fixed length (vectors).

For the Jaccard and cosine distances, the matching of concepts between patients was binary ("all or none"). Semantic similarity between concepts was not considered. Consider a patient A that has the finding *resting tremor*; and a patient B that has the finding *postural tremor*. When calculating the Jaccard distance or the cosine distance, the semantic similarity between *resting tremor* and postural tremor would not contribute to the proximity between these two patients (each metric would value the similarity between *resting tremor* and *postural tremor* as '0'). The semantically augmented

distance metrics behave differently. These augmented distance metrics move patients closer together when patients manifest semantically similar findings, even if they are not exact matches. The augmented cosine distance considers that postural tremor and *resting tremor* have a common immediate ancestor *tremor*. Hence, the *tremor* element of the vectors for patient A and patient B is augmented with a value of 0.5 (see Methods and [19]). This semantic augmentation of the vectors for patients A and B increases their similarity and moves the patients closer together when the cosine distance is calculated (equation 8). The augmented bipartite distance considers that *resting tremor* and *postural tremor* are siblings in the neuro-ontology hierarchy and have a Wu Palmer distance of 0.25 (equation 7); moving patients A and B closer (equations 5 and 6). The augmented cosine distance metric moves the patients closer because *postural tremor* and *resting tremor* have *tremor* as a common ancestor in the neuro-ontology. The augmented bipartite distance metric moves the patients closer because *resting tremor* and *postural tremor* are siblings in the neuro-ontology.

For each of the 382 patients in the dataset (n=382), we calculated the mean patient distance to patients with the same diagnosis and the mean distance to patients with different diagnoses (Figure 2). Within-diagnosis patient distances were lower than between-diagnosis patient distances for all of the metrics (Figure 2). Patients of the same diagnosis should be closer to each other than those with a different diagnosis. Sematic augmentation of the distance metrics makes patients more similar, moves them closer together, and reduces mean patient distances. Augmented cosine and augmented bipartite patient distances were lower than cosine and Jaccard patient distances (Figure 1, Bonferroni post hoc test, $p < .05$). For each patient, the difference between its mean distance to other patients with the same diagnosis and its mean distance to other patients with different diagnosis (Figure 2) is important because it is this difference between within-diagnosis and between-diagnosis distances that contributes to the ability of clustering and classification algorithms to use distances to cluster or classify patients by patient distance successfully [62-63]. The difference between mean within-diagnosis distance and mean-between diagnosis distance differed by metric ($df=3$, $F=49$, $p <.001$) with the largest differences found with the cosine and

augmented cosine metrics and the smaller differences found with the augmented bipartite and Jaccard metrics (Bonferroni post hoc test, $p < .05$).

Classification and clustering.

We evaluated four different classifiers on four different test groups of patients. We used F1 and classification accuracy (Figures 4 and 5) as measures of classification performance. There were differences in classifier performance, with the logistic regression classifier and the k-nearest neighbor classifier outperforming the naïve Bayes classifier (Figures 4 and 5). In retrospect, the selection of the naïve Bayes classifier was ill-suited for this study since this classifier assumes feature independence (not likely to hold among neurological patients) and is oriented towards using probabilities rather than distances for classification. Importantly, we found no effect on classification performance related to the distance metric. Classification performance did vary by test group (Figures 6 and 7). Post hoc testing showed that the classification performance was better for the cranial nerve test group. A likely explanation for the better classification performance with the cranial nerve group is that members of this group (Table 2) had tighter within diagnosis inter-patient distances (i.e., less variability in presentation). As illustrated in Figure 3, the diagnoses of the cranial nerve test group (TN, MNR, RH, On, BEL, BPV, THD, and AN) are primarily on the left-hand side of the x-axis, and they have lower mean intra-diagnosis variability in their clinical presentations.

We evaluated two different clustering algorithms (agglomerative clustering and k-means clustering) on the four test groups of patients (Table 2). Except for the silhouette score, the clustering performance measures depend on the ground truth diagnosis label derived from the patient case studies. The silhouette score measures cluster quality independent of ground truth. Cluster quality did not differ by cluster algorithm (Figure 8). Cluster quality did not vary by distance metric for either the k-means algorithm or the agglomerative algorithm (Figures 9 and 10). Cluster quality did differ by patient test group with post hoc testing showing that the cranial nerve test group had higher cluster quality than the other test groups (Figure 11). Visual inspection of Figures 12 (cranial

nerve test group) and Figure 13 (weakness test group) show how with an 8-cluster solution, cluster *homogeneity* is higher in the cranial nerve group than the weakness test group. In Figures 12 and 13, each color represents a different ground truth diagnosis label, and each column represents a computed cluster. The better performance on clustering of the cranial nerve group likely reflects the same factors intrinsic to this group of patients that led to better classification performance (see above). There is less variability in clinical presentation from patient to patient in this test group, within-diagnosis patient distances are lower (Figure 3), and there is likely less sign and symptom overlap with other diagnoses.

The failure to find an improvement in clustering or classification performance with semantically augmented distance measures was somewhat surprising. Others have found improvements in the clustering of patients [13] or classification of documents [19] with semantically augmented distance metrics. However, Melton et al. [16] did not find improved concordance with domain experts when inter-patient distance calculations were augmented with concept semantic similarity information. Although semantically augmented distance metrics move patients closer (Figure 1), these smaller inter-patient distances may not translate into improvements in clustering or classification performance unless these smaller distances create a greater gap between mean within-diagnosis distance and mean between-diagnosis. From Figure 2, it seems likely that for patients with a given diagnosis, semantic augmented distance places them closer to other patients with the same diagnosis. The problem is that semantically augmented distances push these patients closer to other patients with a different diagnosis. If the net effect of semantic augmentation is to make each patient closer to patients with the same diagnosis and patients with a different diagnosis, there will be no net gain in the ability to cluster or classify patients by diagnosis. The non-intuitive failure of semantic augmentation to improve classification and clustering performance can be illustrated by returning to the hypothetical patient A with *resting tremor* and the hypothetical patient B with *postural tremor.* If the diagnosis of patient A is Parkinson disease and the diagnosis of patient B is essential tremor (as is likely), then semantically augmented distance metrics will move patient A closer to B. However, since the diagnosis of patient A and

patient B are different, moving patient A closer to patient B will deprecate classification and clustering performance in this case.

Implications for neurological diagnosis.

The accuracy of diagnosis for the 32 neurological diagnoses in this study ranged from 76% to 86% with the k-nearest neighbor classifier (Figure 4). In one study, human experts made neurologic diagnoses at the bedside with an accuracy of 77% [64]. Liu et al. [65] observe "machine learning methods can only be as good as the information in the training set…machine-learning methods should not be able to exceed the performance of extremely careful and experienced clinicians…." Machine learning can offer insights into which diseases are more variable in presentation than others (Figure 3) and which diagnostic problems are more challenging to solve than others (Figure 6). Furthermore, machine learning may offer improvements in patient matching strategies for large repositories of archetypal disease profiles such as the Online Mendelian Inheritance in Man [4-5, 12].

Limitations.

One limitation of this study is that we did not consider the severity of deficits, such as weakness or ataxia. When deficits were present, they were binarized as either present or absent and not graded in severity. Another limitation is that some of the diagnosis classes were narrower than others. Although some of the diagnosis classes were specific (Huntington disease, Alzheimer disease, and Parkinson disease), others were more general, such as polyneuropathy, myopathy, and meningitis. This decision to use more general categories for some diagnosis classes reflects the reality that signs and symptoms alone are unlikely to distinguish specific causes of meningitis, polyneuropathy, or myopathy without additional ancillary testing. Another limitation is that we did not compare the computed patient distances to expert opinion for any of the distance metrics. The validity of the results would be improved by a larger dataset of patients, preferably in the thousands rather than in the hundreds. A further limitation of the study is that we utilized published cases from the textbooks of neurology rather than de-identified patient records from electronic medical records. We used manual

abstraction of concepts from case histories instead of natural language processing (NLP) [66-69]. We chose manual abstraction rather than NLP because we wanted to carefully curate a database of test patients with minimal coding errors, and our initial experience with MetaMap indicated that extensive post-processing was needed to ensure accuracy. Future advances in NLP could make the conversion of signs and symptoms in electronic health records to machine-readable codes more accurate and efficient. Inter-rater reliability for abstracting clinical cases into UMLS codes or SNOMED CT codes is another concern [20-21].

Conclusions

Neurological signs and symptoms from case histories can be represented as UMLS concepts from a neuro-ontology. We examined four different distance metrics for the calculation of inter-patient distances. All of the distance metrics provided useful patient distances that could be utilized by machine learning classification and clustering algorithms. Semantically augmented metrics that used the semantic similarity between neurological concepts to calculate patient distances yielded lower patient distances that more traditional distance metrics without semantic augmentation. When each of the four distance metrics was tested on four classifiers and two clustering algorithms, all distance metrics performed similarly without a discernible improvement due to semantic augmentation. Further work is needed to determine the utility of semantically augmenting patient distance metrics with inter-concept distances.

Abbreviations

CUI: UMLS concept unique identifier

UMLS: Unified Medical Language System

SNOMED CT is a registered name of SNOMED International

NLP: Natural language processing

HPO: Human Phenotype Ontology

OMIM: Online Mendelian Inheritance in Man

Declarations

Ethics approval and consent to participate: The Institutional Review Board of the University of Illinois at Chicago approved this work. No consent to participate was required for this work.

Consent to Publish: Not applicable.

Data Availability:

Neurology cases are available at http://dx.doi.org/10.17632/z3d6hwrdmh.2

Inter-concept distances are available at http://dx.doi.org/10.17632/svrx3wgcn4.3

Inter-patient distances are available at http://dx.doi.org/10.17632/svrx3wgcn4.4

Competing Interests: None to report.

Funding:

Partial support for this research was received from the Missouri University of Science and Technology Intelligent Systems Center, the Mary K. Finley Missouri Endowment, the National Science Foundation, the Lifelong Learning Machines program from DARPA/Microsystems Technology Office, and the Army Research Laboratory (ARL); and it was accomplished under Cooperative Agreement Number W911NF-18-2-0260. The research was also sponsored by the Leonard Wood Institute in cooperation with the ARL and was accomplished under Cooperative Agreement Number W911 NF-14-2-0034. The views and conclusions contained in this document are those of the authors. They should not be interpreted as representing the official policies, either expressed or implied, of the Leonard Wood Institute, the ARL, or the United States Government. The United States Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Authors' contributions: Research design by DBH. Data collection by SUB, DBH, and JK. Data analysis by all. Manuscript writing and editing by all.

Acknowledgments: We thank Professor Hal Blumenfeld for permission to reproduce details of the Parkinson case in Table 1 [47].

References

[1] Campbell WW. Diagnosis and localization of neurologic disease, Chapter 53. In Dejong's The neurologic examination. 7th edition. Lippincott Williams and Wilkins, Philadelphia, 2013, pp 769-795.

[2] Beaulieu-Jones B, Finlayson SG, Chivers C, Chen I, McDermott M, Kandola J, Dalca AV, Trends and Focus of Machine Learning Applications for Health Research. (2019) 2: 1–12. doi:10.1001/jamanetworkopen.2019.14051.

[3] Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: A systematic review, J. Biomed. Inform. (2018) 83: 87–96. doi:10.1016/j.jbi.2018.06.001.

[4] Xue H, Peng J, Shang X. Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO, BMC Syst. Biol. (2019) 13: 1–12. doi:10.1186/s12918-019-0697-8.

[5] Peng J, Xue H, Shao Y, Shang X, Wang Y, J. Chen J. Measuring phenotype semantic similarity using Human Phenotype Ontology, Proc. 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016. (2017) 763–766. doi:10.1109/BIBM.2016.7822617.

[6] Pai S, Bader GD. Patient Similarity Networks for Precision Medicine, J. Mol. Biol. (2018) 430: 2924–2938. doi:10.1016/j.jmb.2018.05.037.

[7] Yang S, Stansbury LG, Rock P, Scalea T, Hu PF., Linking Big Data and Prediction Strategies: Tools, Pitfalls, and Lessons Learned, Crit. Care Med. (2019) 47: 840–848. doi:10.1097/CCM.0000000000003739.

[8] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential, Heal. Inf. Sci. Syst. (2014) 2: 1–10. doi:10.1186/2047-2501-2-3.

[9] Deng Y, Gao L, Wang B, Guo X. HPOSim: An r package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology, PLoS One. (2015) 10: 1–12. doi:10.1371/journal.pone.0115692.

[10] Su S, Zhang L, Liu J. An effective method to measure disease similarity using gene and phenotype associations, Front. Genet. (2019) 10: 1–8. doi:10.3389/fgene.2019.00466.

[11] Alanazi HO, Abdullah AH, Qureshi KN. A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care, J. Med. Syst. (2017) 41. doi:10.1007/s10916-017-0715-6.

[12] Köhler S, Schulz MH, Krawitz P, Bauer S, et al. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies, Am. J. Hum. Genet. (2009) 85: 457–464. doi:10.1016/j.ajhg.2009.09.003.

[13] Girardi D, Wartner S, Halmerbauer G, Ehrenmüller M, Kosorus H, Dreiseitl S. Using concept hierarchies to improve calculation of patient similarity, J. Biomed. Inform. (2016) 63: 66–73. doi:10.1016/j.jbi.2016.07.021.

[14] Jia Z, Lu X, Duan H, Li H. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity, BMC Med. Inform. Decis. Mak. (2019) 19: 1–11. doi:10.1186/s12911-019-0807-y.

[15] Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Med Inform*. (2017) 5(1):e7. Published 2017 Mar 3. doi:10.2196/medinform.6730

[16] Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships, J. Biomed. Inform. (2006) 39: 697–705. doi:10.1016/j.jbi.2006.01.004.

[17] Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, Schijvenaars B, Skupin A, Ma N, Börner K. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches, PLoS One. 6 (2011). doi:10.1371/journal.pone.0018029.

[18] L.J. Garcia Castro LJ, R. Berlanga R, A. Garcia A, In the pursuit of a semantic similarity metric based on UMLS annotations for articles in PubMed Central Open Access, J. Biomed. Inform. (2015) 57: 204–218. doi:10.1016/j.jbi.2015.07.015.

[19] Mabotuwana T, Lee MC, Cohen, Solal EV. An ontology-based similarity measure for biomedical data-application to radiology reports. *J Biomed Inform*. (2013) 46(5):857–868. doi:10.1016/j.jbi.2013.06.013

[20] Andrews JE, Richesson RL, Krischer J. Variation of SNOMED CT coding of clinical research concepts among coding experts. J Am Med Inform Assoc. (2007) Jul-Aug;14(4):497-506

[21] Chiang MF, Hwang JC, Yu AC, Casper DS, Cimino JJ, Starren J. Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers, AMIA Annu Symp Proc. (2006) 131–135.

[22] Bhattacharyya SB. Introduction to SNOMED CT. Springer, Singapore, 2016.

[23] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research. (2004) 32, Issue suppl_1, Pages D267–D270, https://doi.org/10.1093/nar/gkh061.

[24] Hier DB, Brint SU. A Neuro-ontology for the neurological examination. BMC Med Inform Decis Mak (2020) 20: 47. https://doi.org/10.1186/s12911-020-1066-7

[25] Choi SS, Cha SH, Tappert CC. A survey of binary similarity and distance measures, WMSCI 2009 - 13th World Multi-Conference Syst. Cybern. Informatics, Jointly with 15th Int. Conf. Inf. Syst. Anal. Synth. ISAS 2009 - Proc. 3 (2009) 80–85.

[26] Tashkandi A, Wiese I, Wiese L. Efficient In-Database Patient Similarity Analysis for Personalized Medical Decision Support Systems, Big Data Res. (2018) 13: 52–64. doi:10.1016/j.bdr.2018.05.001.

[27] Haase P, Siebes R, van Harmelen F. Peer Selection in Peer-to-Peer Networks with Semantic Topologies. In: Bouzeghoub M., Goble C., Kashyap V., Spaccapietra S. (eds) Semantics of a Networked World. Semantics for Grid Databases. ICSNW 2004. Lecture Notes in Computer Science. (2004) vol 3226. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30145-5_7

[28] Rada R, Hafedh M, Bicknell E, Blettner M. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man. and Cybernetics. (1989) 19(1): 17-30.

[29] Wu Z, Palmer M. Verb semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, (1994) pp 133-138.

[30] Leacock C, Chodorow M. Combining Local Context and WordNet Similarity for Word Sense Identification, WordNet. (1998) doi:10.7551/mitpress/7287.003.0018.

[31] Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. (1995) http://arxiv.org/abs/cmp-lg/9511007.

[32] Jiang JJ, Conrath DW. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of International Conference Research on Computational Linguistics (ROCLING X). (1997) Taiwan, pp 19-33, https://www.aclweb.org/anthology/O97-1002.

[33] Lin D. An Information-Theoretic Definition of Similarity, ICML 1998 Proceedings of the Fifteenth International Conference on Machine Learning. (1998) Pages 296-304, July 24 – 27, 1998.

[34] Lee W, Shah N, Sundlass K, Musen M. Comparison of Ontology-based Semantic-Similarity Measures. Medical College of Wisconsin, Milwaukee, WI, Symp. A Q. J. Mod. Foreign Lit. (2008) 384–388.

[35] McInnes BT, Pedersen T. Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs J. Biomed. Inform. (2015) 54: 329–336. doi:10.1016/j.jbi.2014.11.014.

[36] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS, J. Biomed. Inform. (2004) 37: 77–85. doi:10.1016/j.jbi.2004.02.001.

[37] Al-Mubaid H, Nguyen HA, A cluster-based approach for semantic similarity in the biomedical domain, Annu. Int. Conf. IEEE Eng. Med. Biol. Proc. (2006) 2713–2717.

[38] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG, Measures of semantic similarity and relatedness in the biomedical domain, J. Biomed. Inform. (2007) 40: 288–299. doi:10.1016/j.jbi.2006.06.004.

[39] The MathWorks Inc. What is machine learning?, Retrieved at https://www.mathworks.com/discovery/machine-learning.html.

[40] The Mathworks Inc. Supervised learning workflows and algorithms. Retrieved at https://www.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html

[41] The Mathworks Inc. Unsupervised learning. Retrieved at https://www.mathworks.com/discovery/unsupervised-learning.html

[42] Al-Jabery KK, Obafemi-Ajayi T, Olbricht GR, Wunsch II DC (editors). Computational Learning Approaches to Data Analytics in Biomedical Applications, Academic Press. (2020). https://doi.org/10.1016/B978-0-12-814482-4.05001-4.

[43] Rosenberg A, Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure, EMNLP-CoNLL 2007 - Proc. 2007 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn. (2007) 410–420.

[44] Rand WW. Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. (1971) 66: 846–850. doi:10.1080/01621459.1971.10482356.

[45] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. (1987) 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

[46] Kellar SP, Kelvin EA. Munro's Statistical Methods for Healthcare Research. 6th Edition. Wolters Kluwer, Philadelphia, 2013.

[47] Blumenfeld H. Neuroanatomy through Clinical Cases, 2nd Edition. Sinauer Associates, Sunderland, MA, 2010.

[48] Macleod M, Simpson M, Pal S. Clinical Cases Uncovered: Neurology. Wiley-Blackwell, West Sussex UK, 2011.

[49] Noseworthy JH. Fifty Neurologic Cases from Mayo Clinic. Oxford University Press, Oxford UK, 2004.

[50] Pendlebury ST, Anslow P, Rothwell PM. Neurological case histories, Oxford University Press, Oxford UK, 2007.

[51] Toy EC, Simpson E, Mancias P, Furr-Stimming EE. Case Files Neurology, 3rd edition. McGraw-Hill, New York, 2018.

[52] Waxman SG. Clinical Neuroanatomy. 28th Edition. McGraw Hill, New York, 2017.

[53] Hauser SL, Levitt LP, Weiner HL. Case Studies in Neurology for the House Officer. Williams and Wilkins, Baltimore, 1986.

[54] Liveson JA, Spielholz N. Peripheral neurology: case studies in electrodiagnosis. FA Davis Company, Philadelphia, 1979.

[55] Gauthier SG, Rosa-Netto P. Case studies in dementia. Cambridge University Press, Cambridge UK, 2011.

[56] Erro R, Stamelou M, Bhatia K. Case studies in movement disorders. Cambridge University Press, Cambridge UK, 2017.

[57] Solomon T, Michael BD, Miller A, Kneen R. Case studies in neurological infections of adults and children. Cambridge University Press, Cambridge UK, 2019.

[58] Howard J, Singh A. Neurology image-based clinical review. Demos Publishing, New York, 2017.

[59] Pedregosa F, Varoquaux G, Gramfort A, Michel V et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. (2011). 12: 2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html

[60] Jaccard P. The Distribution of the flora in the alpine zone, New Phytologist, (1912) 11: 37–50, doi:10.1111/j.1469-8137.1912.tb05611.x

[61] Ward JH. Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. (1963) 58: 236–244. doi:10.1080/01621459.1963.10500845.

[62] Xu R, Wunsch DC II. Clustering. Wiley-IEEE Press, 2008.

[63] Xu R, Wunsch DC II. Clustering algorithms in biomedical research: A review, IEEE Reviews in Biomedical Engineering. (2010) 3: 120–154.

[64] Chimowitz, MI, Logigian EL, Caplan LR. The accuracy of bedside neurological diagnoses, Ann. Neurol. (1990) 28: 78–85. doi:10.1002/ana.410280114.

[65] Liu Y, Chen PHC, Krause J, Peng L. How to read articles that use machine learning: Users' guides to the medical literature, JAMA - J. Am. Med. Assoc. (2019) 322: 1806–1816. doi:10.1001/jama.2019.16489.

[66] Aronson AR, Lang FM, An overview of MetaMap: Historical perspective and recent advances. J. Am. Med. Informatics Assoc. (2010) 17: 229–236. doi:10.1136/jamia.2009.002733.

[67] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES):

Architecture, component evaluation and applications. J. Am. Med. Informatics Assoc. (2010) 17: 507–513. doi:10.1136/jamia.2009.001560.

[68] Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review, J. Biomed. Inform. (2017) 73: 14–29. doi:10.1016/j.jbi.2017.07.012.

[69] Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes, BMC Med. Inform. Decis. Mak. (2018) 18: 74 doi:10.1186/s12911-018-0654-2.

[70] Jana N, Barik S, Arora N. Current use of medical eponyms--a need for global uniformity in scientific publications. BMC Med Res Methodol. (2009) 9:18. Published 2009 Mar 9. doi:10.1186/1471-2288-9-18

Figures

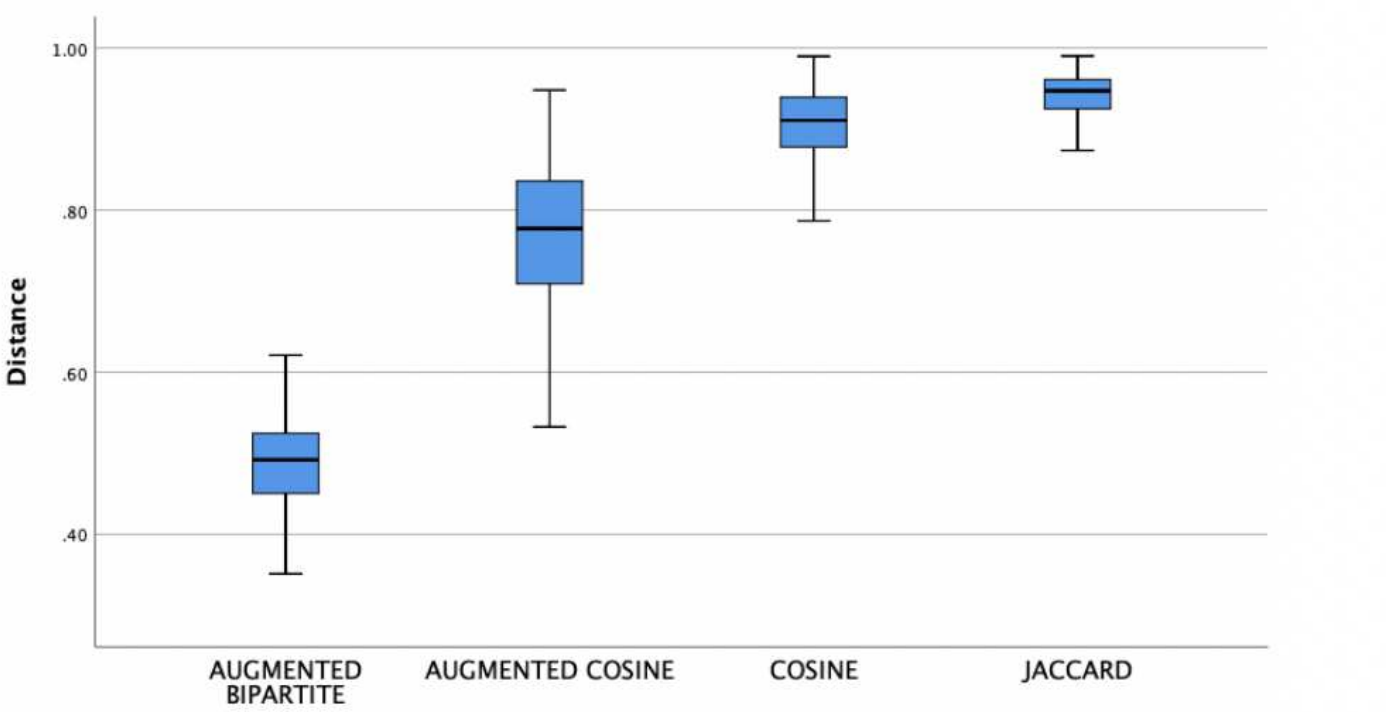

Figure 1

Box-plots inter-patient distances by metric. Means differ by distance metric, (one-way ANOVA, df= 3, F=5820, p <.001). All of the means differed by Bonferroni post hoc test (p <.05) with the Jaccard distance the largest and the augmented bipartite the smallest.

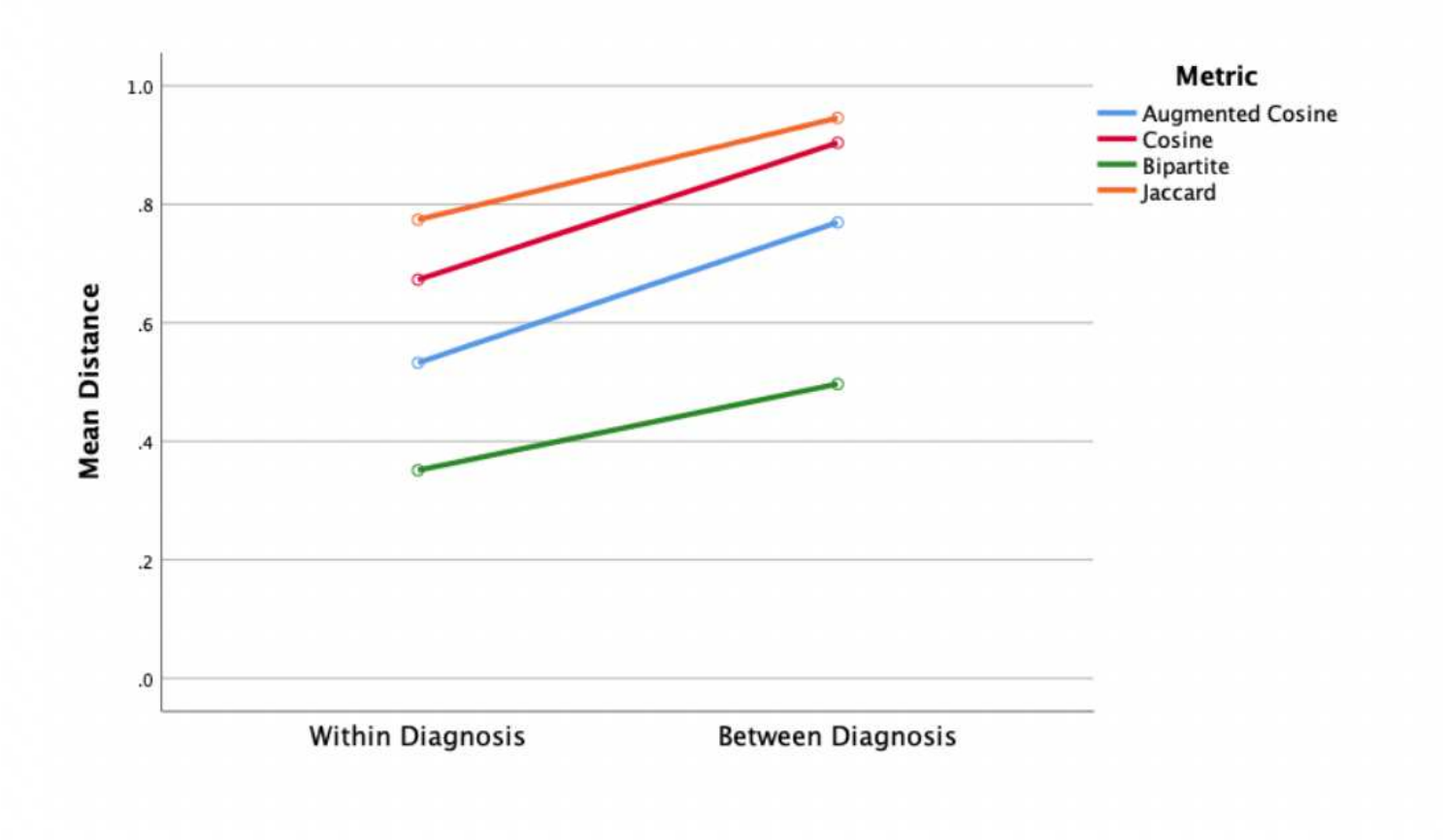

Figure 2

Mean within-diagnosis distance compared to mean between-diagnosis. The within-diagnosis means offer information on patient-to-patient variability within a diagnosis; between-diagnosis means offers information on the degree of separation between patients with one diagnosis from patients of another diagnosis. Mean inter-patient distances were highest for cosine and Jaccard metrics, lowest for augmented bipartite and augmented cosine metrics (post hoc Bonferroni test, $p < .05$). Within-diagnosis mean distances are lower than between-diagnosis mean distances for all metrics (post hoc Bonferroni test, $p < .05$).

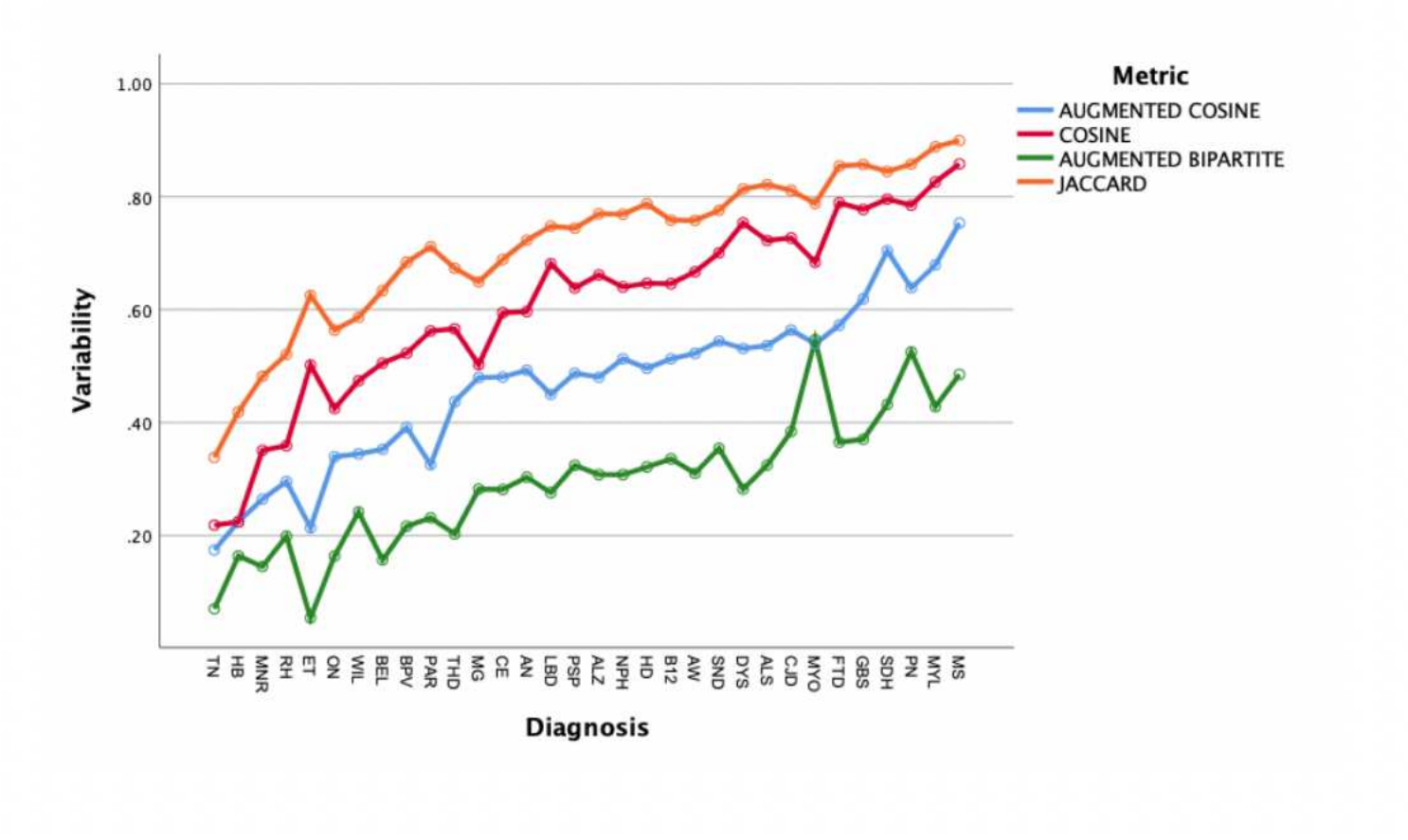

Figure 3

Mean within diagnosis distance by diagnosis in ascending order. Greater within-diagnosis mean patient distance suggests greater variability of clinical presentation within a diagnosis. Diagnoses that are most variable in clinical presentation are to the right of the x-axis. Within-diagnosis mean patient distances vary by diagnosis (two-way ANOVA, df =31, $p < .05$).

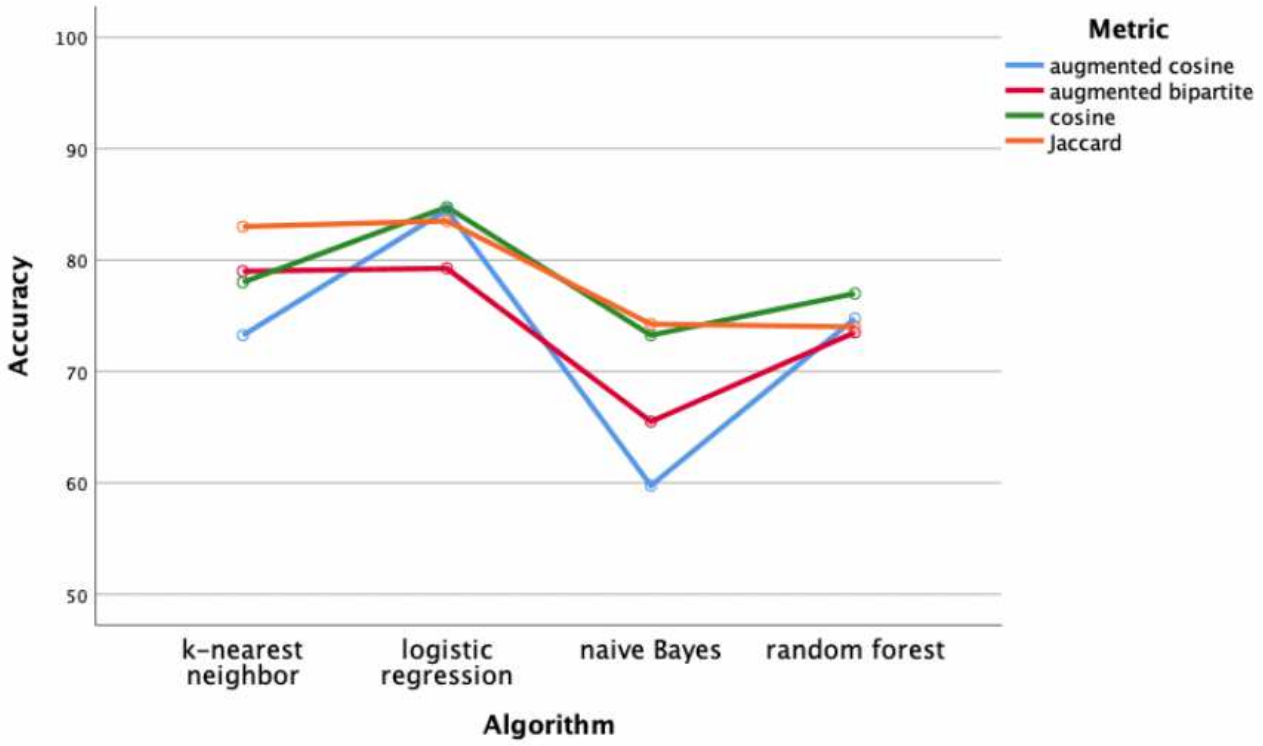

Figure 4

Performance of classifiers by distance metric assessed by classification accuracy. Classification performance on classifiers did not vary by distance metric ($p > .05$). The k-nearest neighbor and logistic regression classifiers outperformed the naïve Bayes classifier (Bonferroni post hoc test, $p < .05$)

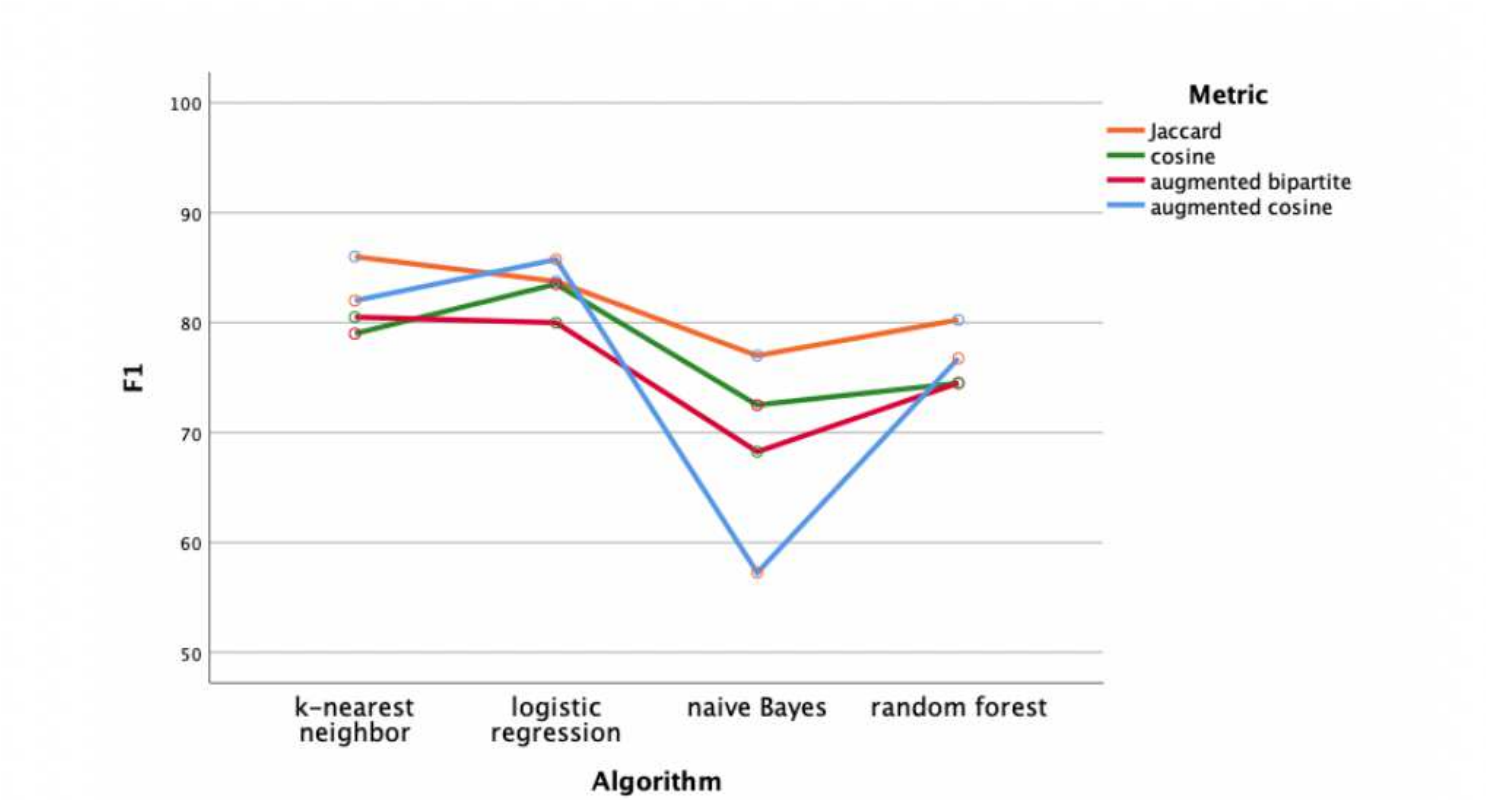

Figure 5

Performance of classifiers by distance metric assessed by balanced F1. Balanced F1 did not vary by distance metric (two-way ANOVA, df =3, p >.05). Naïve Bayes underperformed the k-nearest neighbor and logistic regression classifiers (p <.05).

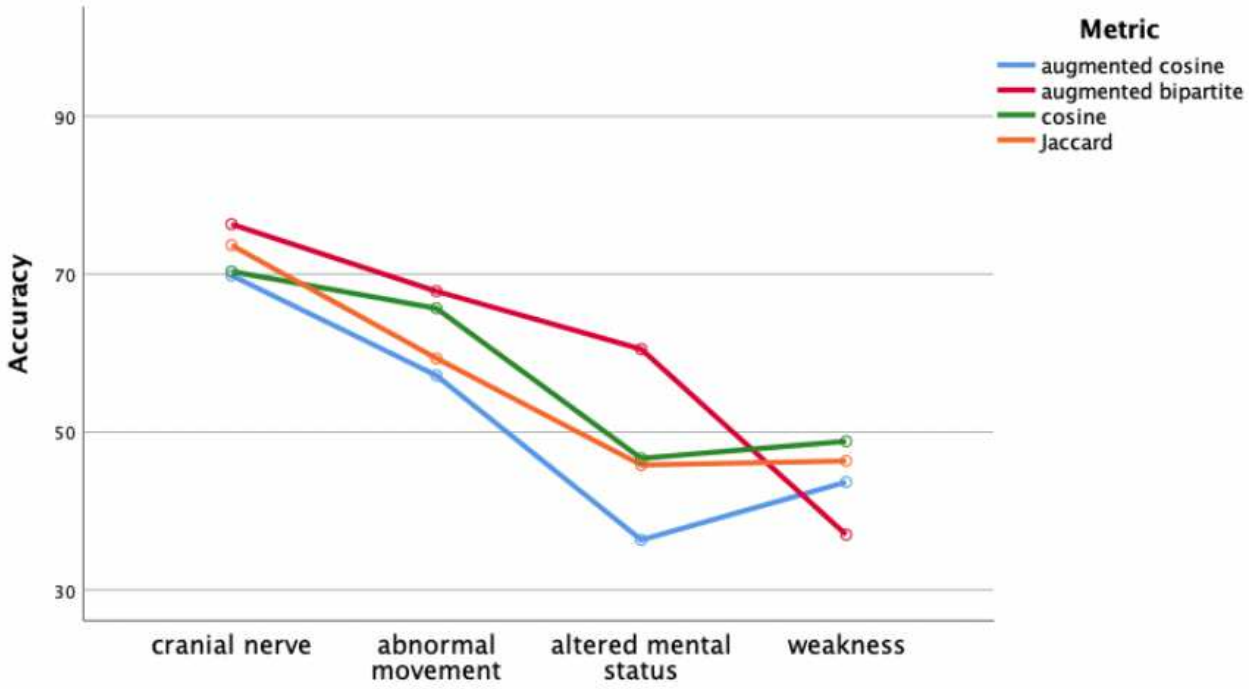

Figure 6

Mean performance of all classifiers by test group assessed by classification accuracy. Classification accuracy did not vary by distance metric (two-way ANOVA, df =3, $p > .05$). Classification accuracy was higher for the cranial nerve group than the other diagnosis groups (Two-way ANOVA, df=3, $p < .01$, post hoc Bonferroni test, $p < .05$).

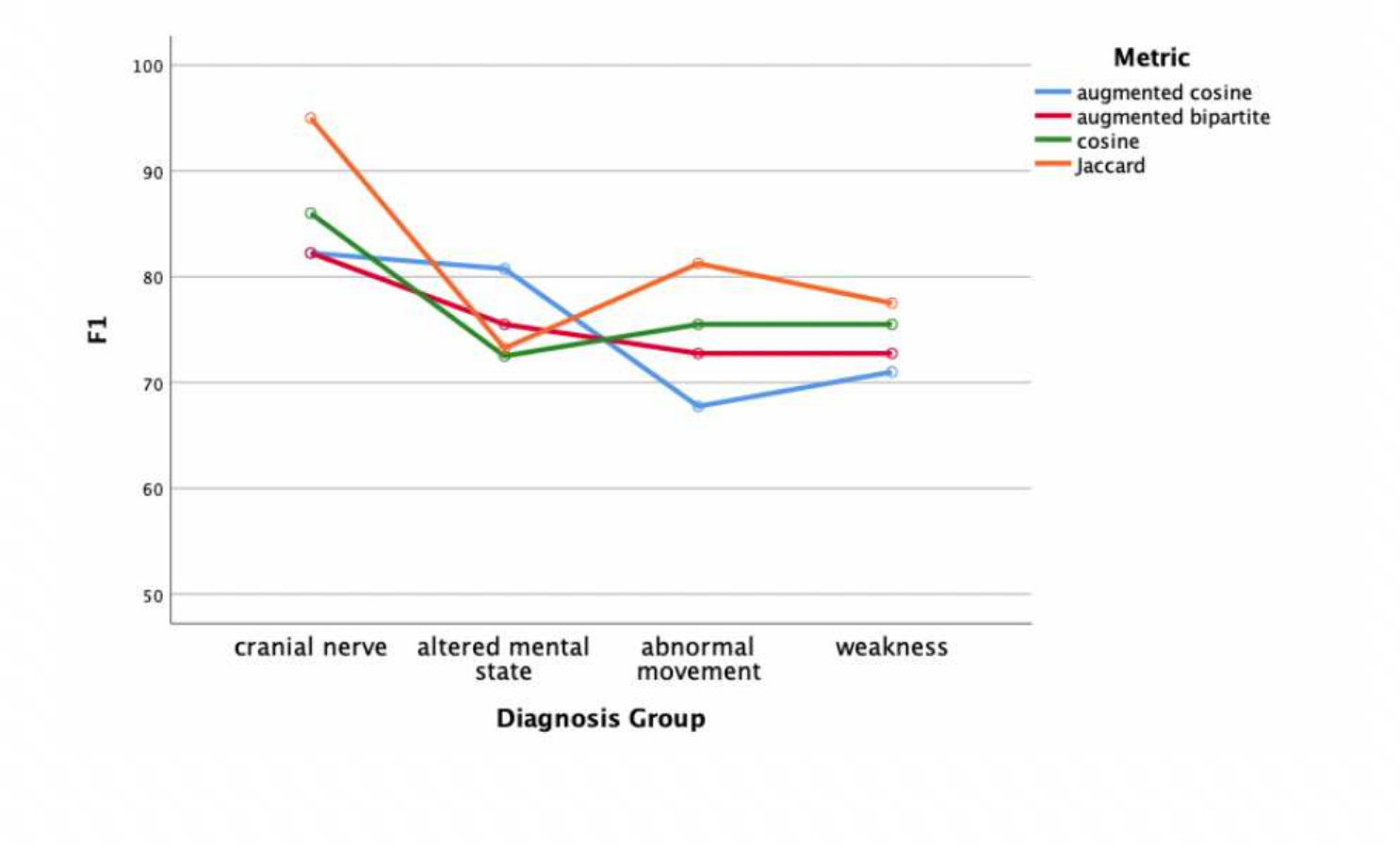

Figure 7

Mean performance of all classifiers assessed by F1 by test group and distance metric. F1 did not vary by distance metric (Two-way ANOVA, df=3, p >.05). F1 varied significantly by diagnosis group (df=3, p <.001, F1 was higher for the cranial nerve test group, p < .05, post hoc Bonferroni test).

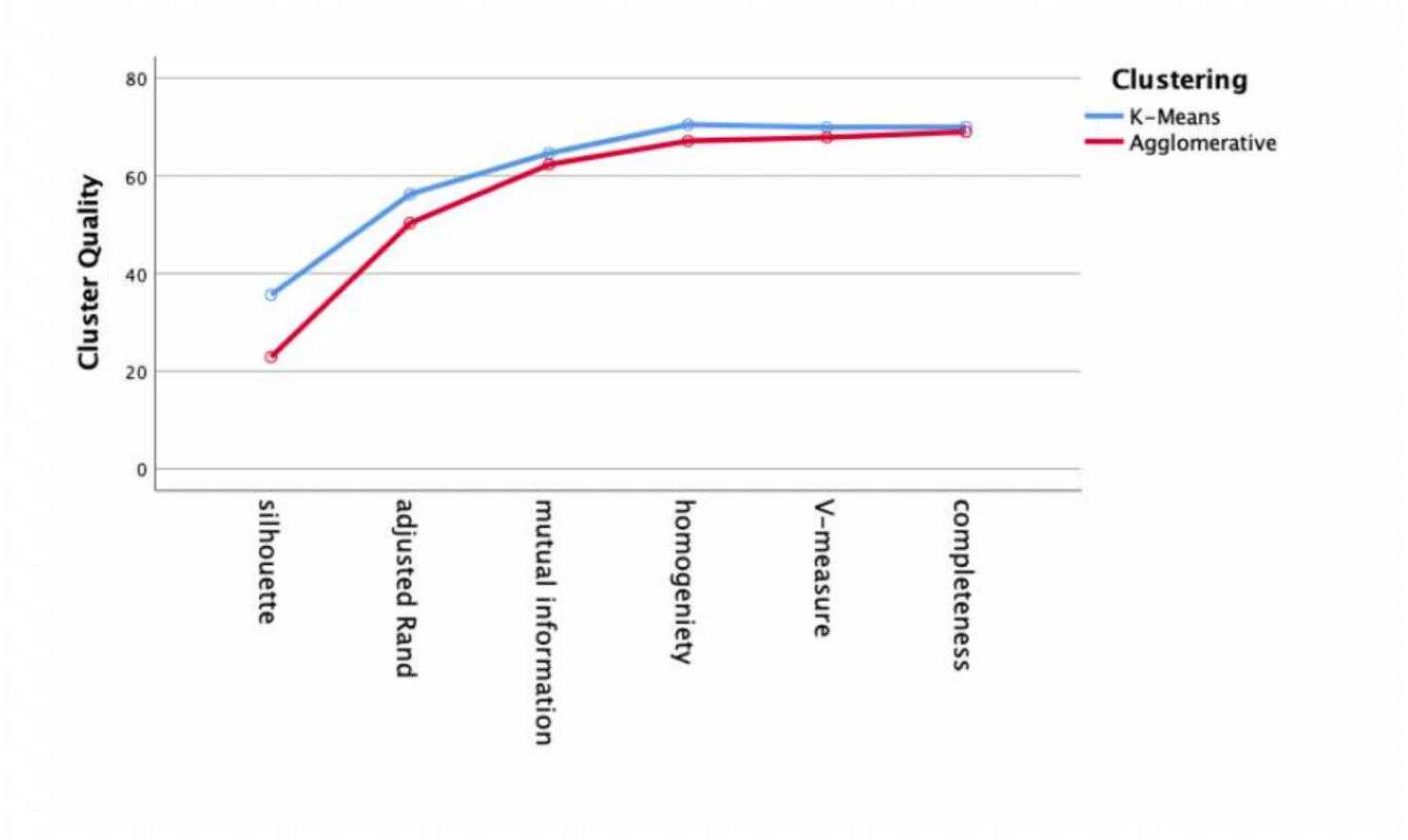

Figure 8

Cluster quality for all test groups comparing k-means to agglomerative clustering (all distance metrics). Cluster quality did not differ by clustering algorithm (two-way ANOVA, df =1, p >.05).

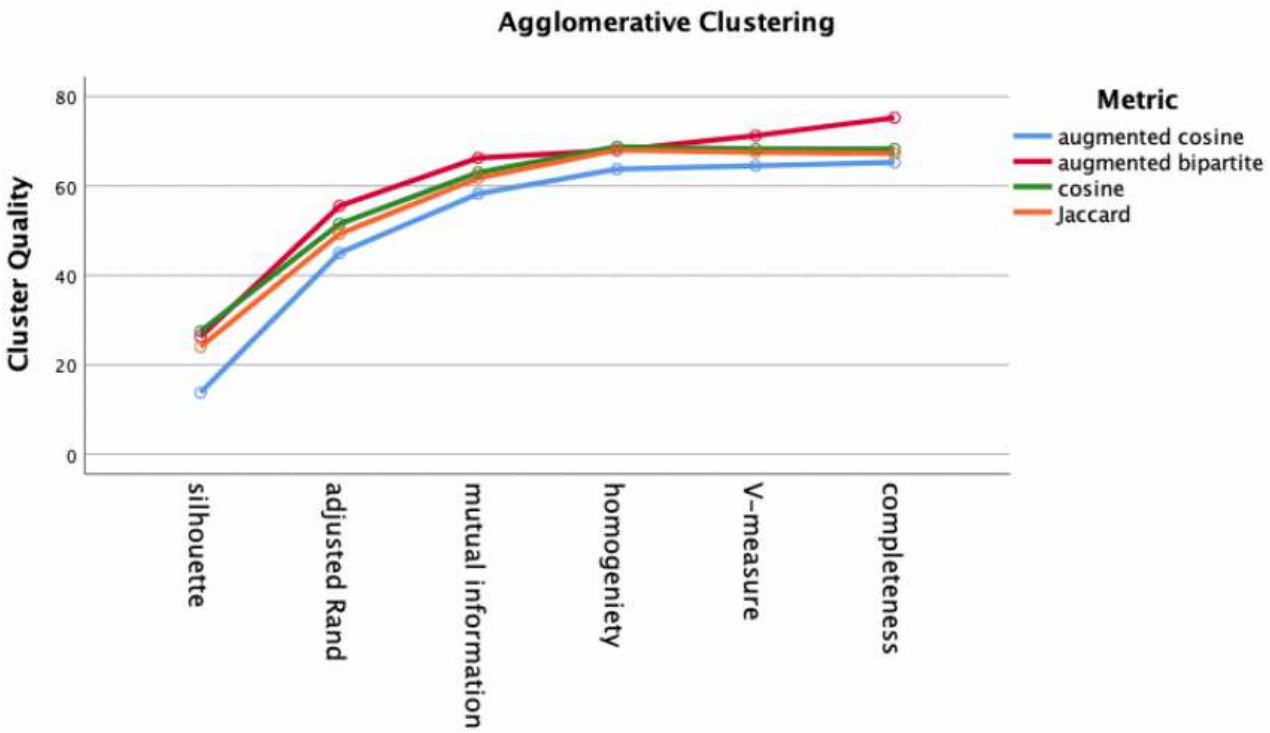

Figure 9

Cluster quality for agglomerative clustering by distance metric. Cluster quality did not differ by distance metric (two-way ANOVA, df =3, p > .05).

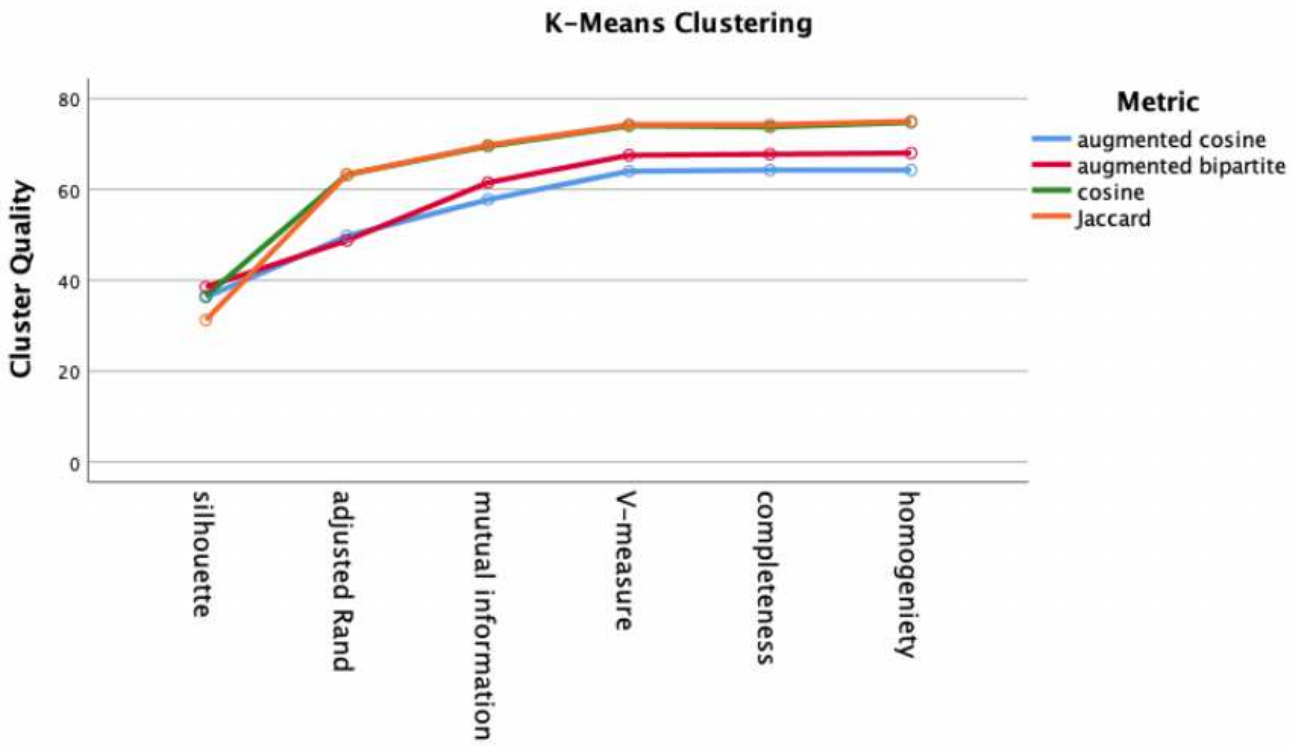

Figure 10

Cluster quality of k-means clustering by distance metric. Cluster quality did not differ by distance metric (two-way ANOVA, df=3, $p > .05$).

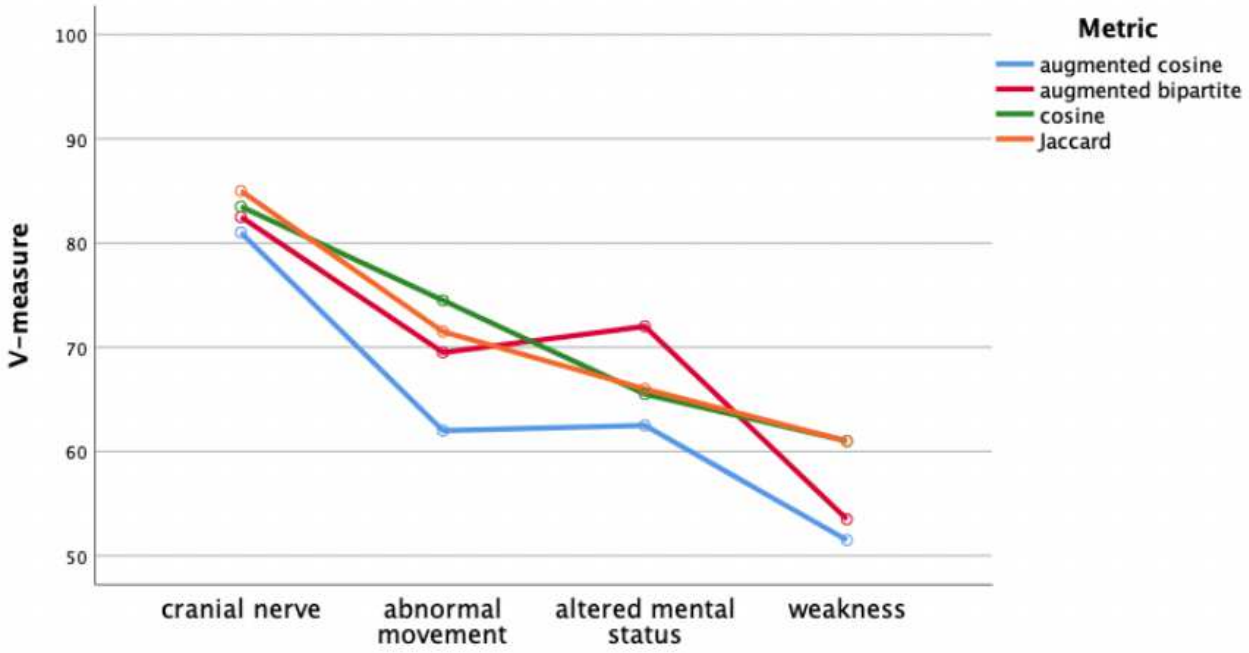

Figure 11

Cluster quality assessed by V-measure by test group and distance metric. Cluster quality did not vary by distance metric (df =3, p > .05). V-measures varied by diagnosis group (two-way ANOVA, df =3, p <.001; post hoc Bonferroni testing showed cranial nerve group to have higher cluster quality, p <.05).

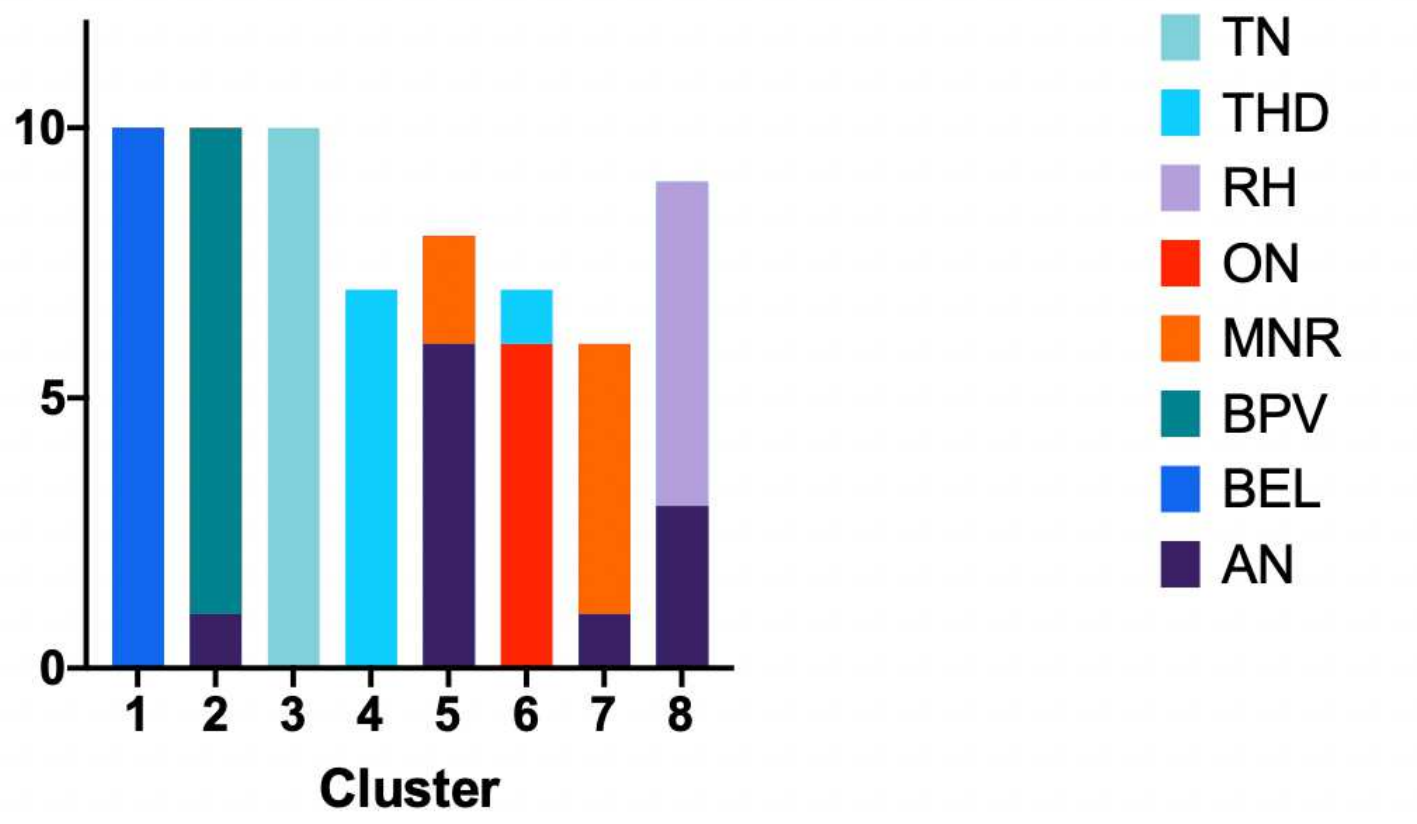

Figure 12

Distribution of ground truth diagnoses by cluster for the cranial nerve test group. K-means clustering with Jaccard distance metric. Each color represents a different ground truth diagnosis. Each column represents a different computed cluster. Homogeneity for the cranial nerve group is greater than for the weakness group (see Figure 13).

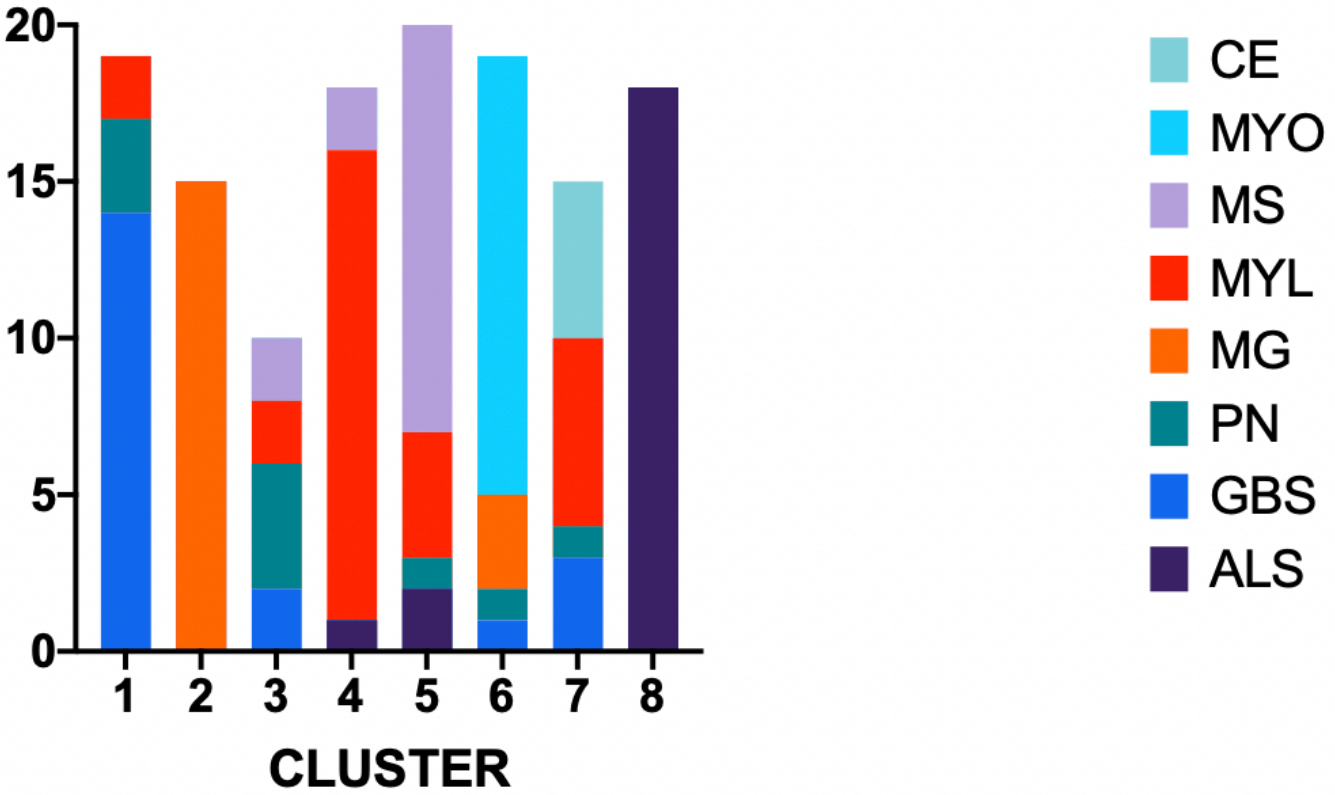

Figure 13

Distribution of ground truth diagnoses by cluster for the weakness test group. K-means clustering with Jaccard distance metric. Each color represents a different ground truth diagnosis. Each column represents a different computed cluster. Homogeneity for the weakness group is less than for the cranial nerve group. (see Figure 12).