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

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

\textsuperscript{a}Department of Neurology and Rehabilitation, University of Illinois at Chicago, Chicago IL

\textsuperscript{b}Department of Internal Medicine, Texas Tech University Health Sciences Center, Lubbock TX

\textsuperscript{c}Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla MO 65401

\textsuperscript{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

-------------------------------

Body and References: 6580 words
Daniel B. Hier  
ORCID 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:
When patient distances are calculated based on phenotype, signs and symptoms are often converted to concepts 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 often dominates 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 signs and symptoms from 382 published neurology cases into sets of concepts with corresponding machine-readable codes. We calculated inter-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 signs and symptoms as the 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:
Using patient diagnoses as labels and patient signs and symptoms as features, we did not find improved classification accuracy or improved cluster quality with semantically augmented distance metrics. Semantic augmentation reduced inter-patient distances but did not improve machine learning performance.
**Background and Related Work**

Patients present with signs (what the physician finds on examination) and symptoms (patient complaints). These signs and symptoms are part of the disease phenotype [1]. Distance metrics play an important role in advancing precision medicine, machine learning, and patient phenotyping [2-11]. In this paper, we focus on metrics that measure the distance between neurological patients based on their signs and symptoms [12]. Signs and symptoms are commonly converted to concepts from hierarchical ontologies before patient distances are calculated. We examine whether semantic augmentation of distance metrics with concept similarities derived from a hierarchical ontology improves the power of distance metrics to classify and cluster neurology patients.

A variety of similarity and distance metrics are available; they have been used to calculate distances between patients [13-16], documents [17-19], and phenotypes [4,5,9,10]. If similarity and distances metrics are normalized to a scale of 0 to 1.0, the distance between A and B is the complement of the similarity.

\[
\text{distance}(A, B) = 1 - \text{similarity} (A, B). \tag{1}
\]

The distance between two patients (or between two documents or between two phenotypes) is different than the distance between two medical concepts. Patients are complex with many dimensions of possible comparison. Inter-patient distances are many-to-many comparisons; inter-concept distances are usually one-to-one comparisons. Metrics that work for concept distances are generally different from metrics to calculate distances between documents, patients, or phenotypes. 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 (signs and symptoms) is challenging. First, neurological symptoms and neurological
signs are recorded in the electronic health record 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 signs and symptoms may vary from patient to patient. Some patients may have as few as one or two signs, while other more complex patients may have as many as 10 or 20 different signs. Fourth, converting unstructured text from electronic health records 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 basis for capturing the signs and symptoms of the neurological examination [24].

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

$$
\text{Sim} (A, B) = \frac{1}{|A|} \times \sum_{a \in A} \max_{b \in B} (\text{sim}(a, b)).
$$ (2)
Bipartite similarity metrics resembling equation 2 have been implemented for the gene-gene and phenotype-phenotype distances in the Human Phenotype Ontology (HPO) [9-10]. Melton et al. [16] adopted a similar metric to measure inter-patient distances.

Medical ontologies such as SNOMED CT and the UMLS Metathesaurus make it possible to calculate distances between concepts utilizing their positions in the hierarchy [28-37]. It has been suggested that distance metrics for both sets and vectors can be augmented by considering the similarity between concepts [13, 14, 19]. Semantically augmented bipartite metrics for patient distance have been suggested by Girardi et al. [13], Melton et al. [18], and Jia et al. [14]. Mabotuwana [19] examined document similarity using a cosine distance metric after converting document concepts to a binarized vector. They found improved classification performance for radiological documents when the distance metric was enhanced semantically with ancestors of the index concepts derived from the SNOMED CT ontology.

Distance metrics are frequently used for clustering and classification of patients. Since the performance of machine learning clustering and classification algorithms can be assessed objectively, we 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 and two clustering algorithms across four different distance metrics. Two of the distance metrics were vector-based and two were set-based; one vector-based metric and one set-based metric was semantically augmented with information about concept similarity. We have tested whether semantic augmentation of the distance metrics improves the performance during classification or clustering.

**Methods**

*Case abstraction.*

We created a dataset of 382 neurological patients based on convenience sampling [37] of published teaching cases [38-49]. We abstracted 2616 signs and symptoms 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 [23]. Table 1 illustrates the case abstraction method for a patient with Parkinson disease.

**Distance metrics**

We implemented four inter-patient distance metrics in python [54]. The Jaccard distance is the complement of the Jaccard similarity [50]. If A and B are the sets of findings from patient A and B, the Jaccard distance \((A, B)\) is shown by equation (3) where \(J_{\text{sim}}\) is the Jaccard similarity.

\[
\text{Jaccard Distance } (A, B) = 1 - J_{\text{sim}}(A, B) = 1 - \frac{A \cap B}{A \cup B}.
\]  

(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 \(a \in A\) and \(b \in B\), the augmented bipartite distance is shown by equation (4) and is supported by equations (5), (6), and (7).

\[
\text{bipartite distance } (A, B) = \frac{D(A, B) + D(B, A)}{2}.
\]

(4)

\[
D(A, B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} \text{dist } (a, b).
\]

(5)

\[
D(B, A) = \frac{1}{|B|} \sum_{b \in B} \min_{a \in A} \text{dist } (a, b).
\]

(6)

\[
\text{dist } (a, b) = 1 - \frac{2 \times \text{depth}(LCS)}{\text{depth}(a) + \text{depth}(b)}.
\]

(7)

For equation (7) we used the hierarchical structure of the neuro-ontology and the method of Wu and Palmer [29] to calculate the \(\text{dist } (a, b)\) as the semantic distance between concept \(a\) and concept \(b\). LCS is the lowest common subsumer in the hierarchical ontology for concepts \(a\) and \(b\); \(\text{depth}(a)\) is number of levels from the root concept to concept \(a\); \(\text{depth}(b)\) is the number of levels from the root concept to concept \(b\), and \(\text{depth}(LCS)\) 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 \( n \times n \) 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 \( (1 - \text{cosine similarity}) \) were calculated by standard methods (equation 8). Patients A and 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 \( b_i \) are 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.

\[
\text{cosine distance} \ (A, B) = 1 - \frac{\sum a_i^2 \ast b_i^2}{(\sqrt{\sum a_i^2}) \ast (\sqrt{\sum b_i^2})}.
\] (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 ancestors, \( n=2 \) for the grandparent ancestors, etc. Descendent concepts 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 a \( n \times n \) 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 \( n \times n \) 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*).

**Classification and Clustering**

For the classification tasks (supervised machine learning), we assessed the classifiers’ performance to assign diagnoses based on signs and symptoms correctly. The ground truth labels were the diagnoses from the abstracted patient histories, and the features were the abstracted signs and symptoms that had been converted to UMLS CUI codes. The performance was assessed by a balanced F1 score and classification accuracy [51] by the cross-validation method with ten folds. Naïve Bayes, logistic regression, random forest, and k-nearest neighbor classifiers were compared. We used the Orange 3.25 default parameters for naïve Bayes, logistic regression (regularization = L2), and random forest (number of trees = 10) classifiers. 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.

For the clustering tasks (unsupervised machine learning), we assessed cluster quality. For both the agglomerative clustering algorithm (Ward linkage) and the k-means clustering algorithm, we chose a parameter of *number of clusters* = 8 based on the 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 [52-55].

**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. 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; each test group had 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-patient 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 pair-wise mean comparisons by group and by distance metric were significant (post hoc Bonferroni test, p < .05).

We found 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 pair-wise 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 diagnosis, 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 n x n 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 signs and symptoms: Jaccard distance, cosine distance, augmented cosine distance and augmented bipartite distance. The calculation of the metrics necessitated different patient representations (see Methods). To calculate the Jaccard and augmented bipartite distances, we represented patients as unordered lists of elements of variable length (sets). In order 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,” and 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 1). Within-diagnosis patient distances were lower than between-diagnosis patient distances for all of the metrics (Figure 1). This is expected, patients of the same diagnosis should be closer to each other than those with a different diagnosis. Semantic 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 [56-57]. 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). Importantly, we found no effect on classification performance related to 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). This is 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 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) shows 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.
Three main conclusions can be drawn from these results. First, usable inter-patient distances can be obtained with any of the four distance metrics. Second, semantic augmentation of the distance measures moves patients reduces patient distance (Figure 1). Third, semantic augmentation of the distance metrics did not improve the performance of classification and clustering algorithms.

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 these same semantic 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 to 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.

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, as did Melton et al. [16]. 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 manually abstracted concepts from case histories rather using natural language processing (NLP) [58-61]. 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. In the future, we plan to extend our methods to a larger dataset of patients and utilize signs and symptoms from de-identified patients in electronic health records. Future advances in NLP could make the conversion of signs and symptoms in electronic health records to machine-readable codes more accurate and efficient. Finally, we did not attempt to identify differing phenotypes within a given diagnosis. The identification of varying clinical presentations (differing phenotypes) within a given diagnosis is an emerging area of research in neurology [65-70]. Interrater reliability for abstracting clinical cases into UMLS codes or SNOMED CT codes is another concern [20-21]. Additional work is needed to establish the reliability of the abstraction process.

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 used 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
References

[1] Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP et al. Finding Our Way through Phenotypes, PLoS Biol. 13 (2015). doi:10.1371/journal.pbio.1002033.

[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, 2 (2019) 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. 83 (2018) 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. 13 (2019) 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. 430 (2018) 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. 47 (2019) 840–848. doi:10.1097/CCM.0000000000003739.
[8] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential, Heal. Inf. Sci. Syst. 2 (2014) 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. 10 (2015) 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. 10 (2019) 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. 41 (2017). doi:10.1007/s10916-017-0715-6.

[12] 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.

[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. 39 (2006) 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. 57 (2015) 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. Singapore: Springer, 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](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. 13 (2018) 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](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](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. 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] Kellar SP, Kelvin EA. Munro’s Statistical Methods for Healthcare Research. 6th Edition. Philadelphia: Wolters Kluwer, 2013.

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

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

[40] Noseworthy JH. Fifty Neurologic Cases from Mayo Clinic. Oxford University Press, Oxford UK, 2004.
[41] Pendlebury ST, Anslow P, Rothwell PM. Neurological case histories, Oxford University Press, Oxford UK, 2007.

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

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

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

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

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

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

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

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

[50] 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
[51] 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.

[52] 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.

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

[54] 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

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

[56] Xu R, Wusch DC II. Clustering. Wiley-IEEE Press, 2008.

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

[58] 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.

[59] 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.

[60] 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.

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

[62] Swinnen B, Robberecht W. The phenotypic variability of amyotrophic lateral sclerosis, Nat. Publ. Gr. (2014). doi:10.1038/nrneurol.2014.184.

[63] Logroscino G, Imbimbo BP, Lozupone M, Sardone R, Capozzo R, P. Battista P, Zecca C, et al. Promising therapies for the treatment of frontotemporal dementia clinical phenotypes: from symptomatic to disease-modifying drugs, Expert Opin. Pharmacother. 20 (2019) 1091–1107. doi:10.1080/14656566.2019.1598377.

[64] Ruggiero L, Mele F, Manganelli F, Bruzzese D, Ricci G, Vercelli L, Govi M, Vallarola A, et al. Phenotypic Variability Among Patients With D4Z4 Reduced Allele Facioscapulohumeral Muscular Dystrophy, JAMA Netw. Open. 3 (2020) e204040. doi:10.1001/jamanetworkopen.2020.4040.

[65] Grad LI, Rouleau GA, Ravits J, Cashman NR. Clinical Spectrum of Amyotrophic Lateral Sclerosis (ALS), Cold Spring Harb. Perspect. Med. 7 (2017) 1–16. doi:10.1101/cshperspect.a024117.

[66] M. Wolff, Ä.K.M. Johannesen, Ä.U.B.S. Hedrich, Ä.S. Masnada, G. Rubboli, E. Gardella, G. Lesca, L. Villard, A. Afenjar, S. Chantot-bastaraud, C. Mignot, L. Perrin,
D. Doummar, M.J. Miranda, Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders, (2017) 1316–1336. doi:10.1093/brain/awx054.

[67] Kantarci OH. Phases and phenotypes of multiple sclerosis, Contin. Lifelong Learn. Neurol. 25 (2019) 636–654. doi:10.1212(CON).00000000000000737.

[68] 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

Abbreviations

CUI: UMLS concept unique identifier

UMLS: Unified Medical Language System

SNOMED CT is a registered name of SMOMED International

NLP: Natural language processing
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/svrx3wgcnc4.3

Inter-patient distances are available at http://dx.doi.org/10.17632/svrx3wgcnc4.4

Competing Interests: None to report.

Funding: None to report.

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

Acknowledgements: We thank Professor Hal Blumenfeld for permission to reproduce details of the Parkinson case in Table 1 [38].