Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals

Brian Connolly,[1] Pawel Matykiewicz,[1] K Bretonnel Cohen,[2] Shannon M Standridge,[3] Tracy A Glauser,[3] Dennis J Dlugos,[4] Susan Koh,[5] Eric Tham,[6,7] John Pestian[1]

[1]Cincinnati Children's Hospital Hospital Medical Center, Division of Bioinformatics, Cincinnati, Ohio, USA
[2]University of Colorado School of Medicine, Biomedical Text Mining Group, Computational Bioscience Program, Aurora, Colorado, USA
[3]Cincinnati Children's Hospital Medical Center, Division of Pediatric Neurology, Cincinnati, Ohio, USA
[4]The Children's Hospital of Philadelphia, Division of Child Neurology, Philadelphia, Pennsylvania, USA
[5]University of Colorado/ Children's Hospital Colorado, Division of Pediatric Neurology, Aurora, Colorado, USA
[6]University of Colorado School of Medicine, Department of Pediatrics, Aurora, Colorado, USA
[7]Children's Hospital Colorado, Research Informatics, Aurora, Colorado, USA

Correspondence to
Dr John Pestian, 3333 Burnet Ave, MLC 7024, Cincinnati, OH 45229-3039, USA; john.pestian@cchmc.org

Received 29 December 2013
Revised 14 February 2014
Accepted 6 March 2014
Published Online First 1 April 2014

To cite: Connolly B, Matykiewicz P, Bretonnel Cohen K, *et al*. *J Am Med Inform Assoc* 2014;**21**: 866–870.

ABSTRACT

Objective The constant progress in computational linguistic methods provides amazing opportunities for discovering information in clinical text and enables the clinical scientist to explore novel approaches to care. However, these new approaches need evaluation. We describe an automated system to compare descriptions of epilepsy patients at three different organizations: Cincinnati Children's Hospital, the Children's Hospital Colorado, and the Children's Hospital of Philadelphia. To our knowledge, there have been no similar previous studies.

Materials and methods In this work, a support vector machine (SVM)-based natural language processing (NLP) algorithm is trained to classify epilepsy progress notes as belonging to a patient with a specific type of epilepsy from a particular hospital. The same SVM is then used to classify notes from another hospital. Our null hypothesis is that an NLP algorithm cannot be trained using epilepsy-specific notes from one hospital and subsequently used to classify notes from another hospital better than a random baseline classifier. The hypothesis is tested using epilepsy progress notes from the three hospitals.

Results We are able to reject the null hypothesis at the 95% level. It is also found that classification was improved by including notes from a second hospital in the SVM training sample.

Discussion and conclusion With a reasonably uniform epilepsy vocabulary and an NLP-based algorithm able to use this uniformity to classify epilepsy progress notes across different hospitals, we can pursue automated comparisons of patient conditions, treatments, and diagnoses across different healthcare settings.

BACKGROUND AND SIGNIFICANCE

About 3% of the American population will develop epilepsy by the age of 75.[1] The diagnosis of epilepsy significantly affects a patient's quality of life, and also imposes a considerable medical and economic burden.[2] Despite the high prevalence of epilepsy, there is an epilepsy–treatment gap, defined as the proportion of people with epilepsy who require treatment but do not receive it.[3] Racial, ethnic, and socioeconomic disparities in access to healthcare all contribute to this gap.[4] Recognizing the problem, the American Academy of Neurology introduced eight quality of care measures.[5] It is important for completing these measures that it is determined whether or not there is a degree of uniformity in the clinical vocabulary used by different hospitals—if there is no such uniformity, then we cannot use typical natural language processing (NLP) methods to compare quality measures, as superficial linguistic differences will spuriously suggest quality differences. An additional benefit of this research is the development of an NLP algorithm that can be used to compare clinical notes between different hospitals without the need for a huge amount of training data; such algorithms will become increasingly important in large-scale analyses of healthcare delivery and outcomes.

To our knowledge, no prior statistical analysis has directly quantified similarities and/or differences in the vocabulary of clinical notes across institutions. However, there has been considerable effort in understanding inter- and intrahospital similarities and differences in clinical notes, as well as the medical terminology used in them. For instance, in an effort of evaluate interhospital differences Uzuner *et al*,[6] was able to train a machine learning technique using discharge summaries from one hospital to classify assertions in discharge summaries and radiology reports from others. Fan *et al*[7] evaluated the ability of an NLP algorithm to tag parts of speech in one medical institution after it was trained using notes from a second institution. They found that a machine learning model was able to tag parts of speech in clinical notes from Kaiser Permanente Southern California with 89% accuracy when trained on clinical notes from the University of Pittsburgh Medical Center.[8] Matykiewicz *et al*[9] presented direct intrahospital comparisons of clinical notes which utilized a support vector machine (SVM) and the K-L divergence to determine whether or not there were differences in the n-gram frequencies in clinical notes from patients with intractable and non-intractable epilepsy. The latter analysis is directly related to this work: one could replace notes from patients with intractable and non-intractable epilepsy with epilepsy progress notes from different institutions and thereby directly measure the similarities in their vocabulary. However, such an approach would not address whether or not the linguistic differences among different hospitals were important (one would expect large-scale interhospital differences from the electronic health record template formatting of the notes).

OBJECTIVE

We tackle the problem of quantifying the similarities and differences in epilepsy clinical notes by developing an SVM that uses surface linguistic features, and then testing whether or not the interhospital differences in vocabulary are sufficient to prevent the SVM from correctly classifying an epilepsy progress note as one describing a patient with

generalized (GE), partial (PE), or unclassified epilepsy (UE). PE and GE cover 88% of epilepsy patients.[10] Given the ambiguous nature of UE assigned by a clinician, but also considering its relatively high prevalence, we first classify the notes using PE and GE, only, and then perform a separate analysis including PE, GE, and UE. We further investigate whether the classification can be improved by training the SVM using progress notes from multiple hospitals.

Our approach rests on the assumption that if an SVM is trained on epilepsy progress notes from one hospital and is then able to successfully classify notes from another hospital, then the hospitals share a common epilepsy vocabulary. Supporting evidence for this assumption would suggest a nationwide commonality in epilepsy vocabulary which can be exploited for epilepsy information extraction (eg, annotation and classification of epilepsy progress notes across different hospitals) for quality measures.

METHODS AND MATERIALS

We train an SVM using epilepsy progress notes from one or two hospitals. The SVM classifies the notes based on the frequencies of (strings of) words (n-grams) in the notes. The common vocabulary is therefore strictly defined by those n-grams that are associated with the classifications. However, the classifications are broad enough that reasonable inferences could be made regarding the general epilepsy vocabularies of the hospitals. The SVM is trained to classify each progress note as belonging to a patient with one of three broadly defined categories of epilepsy: PE, GE, and UE. Due to the lack of consensus in their annotation, the epilepsy progress notes are defined by the ICD-9-CM codes assigned to them by their authors with GE defined by 345.00, 345.01, 345.10, 345.11, and 345.2; PE defined by 345.40, 345.41, 345.50, 345.51, 345.70, and 345.71; and UE defined by 345.80, 345.81, 345.90, and 345.91. Note that the codes themselves never occur in the notes, and since the clinicians are not required to use any controlled vocabulary, the text strings associated with the codes most likely never occur in the notes either.

Table 1 summarizes the ICD-9-CM codes and lists the numbers of progress notes available for classification for each hospital. As there are sizable variations in the number of notes between the three epilepsy types, using them all would result in sample-size effects that could be confused with interhospital differences in vocabulary. We therefore fix the training and data sample sizes to 90 documents per hospital per epilepsy classification in the training set, and to 45 documents per hospital per epilepsy classification in the testing data set. The training set is used for two purposes: for cross-validation of the parameter space and for building the optimal classifier. The test set (ie, 'remaining hospital(s)') is withheld until the optimal classifier is built on the full training data.

Table 1 The ICD-9-CM codes associated with each type of epilepsy diagnosis, and the corresponding number of clinical notes from each hospital

Epilepsy classification	ICD-9-CM codes	CCHMC	CHCO	CHOP
Partial epilepsy	345.40, 345.41, 345.50, 345.51, 345.70, 345.71	303	128	269
Generalized epilepsy	345.00, 345.01, 345.10, 345.11, 345.2	99	163	129
Unclassified epilepsy	345.80, 345.81, 345.90, 345.91	200	117	121
Data missing	345.3, 345.60, 345.61	12	25	32

CCHMC, Cincinnati Children's Hospital Medical Center; CHCO, Children's Hospital Colorado; CHOP, Children's Hospital of Philadelphia.

To validate the gold standard in the face of known problems with practitioner-assigned ICD-9-CM codes, a random sample of 24 notes from each category was assembled. Each note was annotated by two physicians, with each physician only coding the notes from the hospital(s) other than their own. This process resulted in a Krippendorff's α of 0.691 (with chance agreement of 1/4), suggesting that the gold standard is of good quality. When we combined the post hoc coding with the coding done by the authors of the notes, Krippendorff's α slightly decreased to 0.626.

The documents are represented by their unigrams, bigrams, and trigrams, which serve as features for the SVM. We found that the inclusion of n-grams with n larger than 3 decreases classification accuracy (the F_1 score described below) during training, probably due to over-fitting. The extraction of n-grams is described in the following section. This is the most basic representation that could be used. An alternative approach would be to use semantic features, rather than surface linguistic features, by running a term extraction engine such as MetaMap, cTAKES, or ConceptMapper, and then classifying based on the extracted semantic concepts. As will be seen, good classification can be obtained with the simpler approach. Furthermore, abstraction of semantic concepts has the effect of making the three hospitals more homogeneous, so the surface linguistic features provide a more stringent evaluation of the hypothesis.

N-gram extraction

We used the electronic health records from the neurology departments of three different hospitals: the Cincinnati Children's Hospital Medical Center (CCHMC), Children's Hospital Colorado (CHCO), and Children's Hospital of Philadelphia (CHOP). The progress notes were required to have been created for an office visit, be over 100 characters in length, and have one of the ICD-9-CM codes listed in table 1. Further, each note had to be signed by an attending clinician, resident, fellow, or nurse practitioner. Lastly, each patient was required to have at least one visit per year between 2009 and 2012 (for a minimum of four visits). Overall, 551, 614, and 433 progress notes from CHOP, CCHMC, and CHCO, respectively, satisfied all of the selection criteria.

The notes were then de-identified using a combination of automatic output from the MITRE Identification Scrubber Tool (MIST)[11] and manual review. After de-identification, the n-gram frequencies were extracted from each note, and all characters in the note were changed to lower case. Age, patient name, location, hospital name, any initials, patient identification numbers, phone numbers, URLs, and miscellaneous protected information such as account numbers and room numbers were replaced with 'AGE,' 'NAME,' 'LOCATION,' 'HOSPITAL,' 'INITIALS,' 'ID,' 'PHONE,' 'URL,' and 'OTHER,' respectively. Non-ASCII and non-alphanumeric characters were then removed, and all numbers were changed to 'NUMB.' All n-grams that occurred less than nine times within the whole data set were removed.

Progress note classification

SVMs are the most commonly used machine learning technique for text classification tasks. They are particularly useful in cases such as that presented in this work, where the training sample is small but feature rich.[12] Further, in Matykiewicz *et al*,[9]

statistical and machine learning methods were compared based on their ability to classify epilepsy progress notes as describing patients with either intractable or non-intractable epilepsy. The SVM proved to be the best classifier, given differences were indeed present between the two classes of notes (an SVM cannot quantify similarity). Both the progress notes and classification scheme described in this work are similar to those in Matykiewicz *et al*.[9]

The SVMs are trained using 90 documents for each of the three epilepsy types, with as many as 23 017 n-grams, and optimized using an F_1 score defined by

$$F_1 = \frac{2t_n^2}{(t_n + f_p)(t_n + f_n)} \quad (1)$$

where t_n is the number of true positives, f_p is the number of false positives, and f_n is the number of false negatives.

N-grams are weighted based on one of two weighting schemes. The schemes are selected using cross-validation methods, among other parameters. Ultimately, the SVM is optimized over the cost regularization parameter (the C parameter), the number of top-ranked n-grams to use for the SVM input (N), and the ranking method and n-gram weighting schemes using the 20-fold cross-validated F_1 score. The cost parameter is optimized over 18 values ranging from 2^{-8} to 2^4, incremented by factors of 2. Parameter N is optimized over 2^5 to 2^{13} n-grams, incremented by factors of $2^{0.5}$.

The n-grams are ranked based on either information gain, information gain ratio, or the Pearson correlation coefficient. Overall, the SVM is optimized over 13 values of the C parameter, 16 values of N, 2 feature weightings, 3 feature rankings, and 20 folds. This translates to an optimization over 1248 points in the parameter space and 24 960 runs of the SVM.

As discussed previously, the UE classification can be ambiguous. We therefore classify GE and PE for three hospitals using training samples from either one or two of the other hospitals. This gives six possible combinations of hospitals. The baseline classifier for these experiments is random class assignment, which yields F_1=50%.

We also perform a second analysis assuming three possible types of epilepsy—PE, GE, and UE. Because SVMs are built for binary classification, three SVMs are trained to classify PE versus not-PE, GE versus not-GE, and UE versus not-UE, with the results being subsequently combined to effectively provide a tertiary classification. The baseline classifier for these experiments is F_1=33%.

RESULTS

Table 2 summarizes the performance of our SVM trained assuming patients are either PE or GE. It shows 20-fold cross-validated F_1's and corresponding SDs for both GE and PE progress notes. The corresponding average F_1's and their SDs from progress notes sampled from the hospitals not in the training set (ie, 'remaining hospitals') are also listed along with the p value significance, which assume a random baseline classification of F_1=50%. The p values show the SVM is capable of classifying PE and GE above baseline, although the p value in the case where the training sample is CCHMC and the F_1 is evaluated on CHOP and CHCO is significantly smaller than in the case when the SVM is trained and evaluated with other training and testing data sets. Note that the F_1's are all above approximately 75% when the SVM is trained on two hospitals. Also, training with two hospitals yields an increase of about 10.4% in F_1. The other effect of adding a second hospital is the decreased gap between training F_1 and testing F_1. The gap 0.871–0.725=0.146 decreases to 0.899–0.829=0.070, yielding a 7.6% improvement. All three effects suggest that two hospitals are enough to make the third one more similar.

The results from our second study, where we include patients with UE, are shown in table 3. The F_1 scores are all above the baseline value of 33%, although somewhat marginally. As before, there is a 10.4% improvement in F_1 when a second hospital is added to the training set and the F_1 gap between the training and testing sets decreases from 0.289 to 0.216, which is an improvement of about 7.3%.

Although the changes in the second study are marginal, they do not contradict our previous conclusions. Most likely the notes from UE patients obscure the classification of GE and PE, as words associated with both would also appear in the UE notes.

DISCUSSION

We have developed an SVM classifier with surface linguistic features that supports the rejection of our null hypothesis (which is that such an algorithm cannot be trained using epilepsy-specific notes from one hospital and then successfully used to classify epilepsy patients from another hospital) with statistical significance. We have therefore established a certain uniformity among epilepsy progress notes from three different institutions: the

Table 2 Results from the classification of partial epilepsy and generalized epilepsy in epilepsy progress notes

Hospital used for training	Average F_1 (training)	F_1 SD (training)	Average F_1 (remaining hospitals)	F_1 SD (remaining hospitals)	p Value from baseline (remaining hospitals)
CCHMC	0.865	0.213	0.691	0.095	0.043
CHOP	0.926	0.149	0.729	0.014	<0.001
CHCO	0.823	0.224	0.754	0.062	<0.001
One-hospital average	0.871	0.195	**0.725**	0.070	0.001
CCHMC and CHOP	0.913	0.100	0.817	0.047	<0.001
CCHMC and CHCO	0.904	0.097	0.807	0.031	<0.001
CHOP and CHCO	0.904	0.097	0.807	0.031	<0.001
Two-hospital average	0.899	0.105	**0.829**	0.047	<0.001

The first column lists the hospital(s) used to optimize the support vector machine. The second and third columns list the 20-fold cross-validated average F_1 and corresponding SDs of the training samples, respectively. The fourth and fifth columns list the average F_1 and corresponding SDs for the remaining hospital(s). The last column shows the p value significance of the result compared to the largest class baseline F_1=0.5. Systematic improvement when two hospitals are used is highlighted in bold, and the sample size is the same when one and two hospitals are used.
CCHMC, Cincinnati Children's Hospital Medical Center; CHCO, Children's Hospital Colorado; CHOP, Children's Hospital of Philadelphia.

Table 3 Results from the classification of PE, GE, and UE in epilepsy progress notes

Hospital used for training (remaining hospitals)	Average F_1 (training)	F_1 SD (training)	Average F_1 (remaining hospitals)	F_1 SD	p Value from baseline (remaining hospitals)
CCHMC	0.647	0.311	0.417	0.147	0.567
CHOP	0.759	0.261	0.372	0.142	0.788
CHCO	0.625	0.327	0.376	0.143	0.763
One hospital	0.677	0.300	**0.388**	0.145	0.704
CCHMC and CHOP	0.730	0.169	0.478	0.097	0.136
CCHMC and CHCO	0.670	0.185	0.574	0.191	0.207
CHOP and CHCO	0.724	0.172	0.424	0.113	0.421
Two hospitals	0.708	0.175	**0.492**	0.153	0.298

The first column lists the hospital(s) used to optimize the support vector machine. The second and third columns list the 20-fold cross-validated average F_1 and corresponding SDs of the training samples, respectively. The fourth and fifth columns list the average F_1 and corresponding SDs for the remaining hospital(s). The last column shows the p value significance of the result compared to the largest class baseline $F_1 \approx 0.333$. Systematic improvement when two hospitals are used is highlighted in bold, and the sample size is the same when one and two hospitals are used.
CCHMC, Cincinnati Children's Hospital Medical Center; CHCO, Children's Hospital Colorado; CHOP, Children's Hospital of Philadelphia; GE, generalized epilepsy; PE, partial epilepsy; UE, unclassified epilepsy.

CCHMC, CHCO, and CHOP. The document/n-gram matrix was built using unigrams, bigrams, and trigrams, and employed for training SVM text classifiers.

We also showed that for a given (fixed) number of progress notes, the classification of patient notes from a third hospital is improved by using notes from two hospitals in the SVM training set. That is, given the choice of increasing the sample size by increasing the number of notes from a single hospital, or broadening the note pool by including notes from another hospital, our results suggest the latter is the better choice for classification. Our results suggest the inclusion of a second hospital may yield an improvement. The case where the training sample is CCHMC progress notes and the model is evaluated on CHOP and CHCO progress notes gives a significance of ~5%, whereas those cases where two hospitals are included in the training set all yield an improvement over baseline that is statistically significant at a p value of <0.01.

We are conscious of certain assumptions and possible biases that are inherent in this analysis; however, we believe they do not invalidate our conclusions. For example, while our selection criteria for the progress notes introduce biases in the number of GE and PE progress notes, they are not relevant as we train and classify the SVM on a fixed number of each. Another factor to consider is that the GE and PE classifications are defined by the ICD-9-CM codes, which are used primarily to encode billing information. Although strictly speaking they cannot be used as proxies for clinical diagnoses, they were assigned by the author of the note (a healthcare provider).

It is worthwhile noting that while technically our method does not directly address whether clinical vocabulary is the same across all three institutions considered, given our results one can reasonably infer that a degree of similarity exists. Natural language, including clinical language, is complex and ambiguous at the level of vocabulary and at the levels of morphology, syntax, semantics, and document structure. Even so, we cannot conclude that the notes are heterogeneous across the three hospitals based on lexical features only. In fact, we know that the different EPIC templates used in the different hospitals introduce differences in the document structure and semantic concepts mentioned in the records. However, our findings are consistent with the idea that even in the face of these additional levels of complexity, surface linguistic features alone do not introduce spurious indicators of differences in quality measures across the three hospitals.

CONCLUSIONS

Our work has established that there is a certain degree of uniformity of epilepsy vocabulary across different hospitals, and has developed an NLP-based machine learning technique to classify and extract information from epilepsy progress notes. This suggests that a limited number of annotated epilepsy progress notes from each hospital might be enough for developing automated extraction of epilepsy quality measures from clinical narratives.

Contributors Deborah Batson (Children's Hospital Colorado); Marianne Chilutti (Children's Hospital of Philadelphia); Robert Faist (Cincinnati Children's Hospital Medical Center); Erin Murphy (Cincinnati Children's Hospital Medical Center); Lesley Rohlfs (Cincinnati Children's Hospital Medical Center); Gregory Schulte (Children's Hospital Colorado); Christine Wilson (Cincinnati Children's Hospital Medical Center).

Contributors BC collaborated with PM on the analysis and presentation of results in this manuscript. KBC and JP conceptualize the project and provided guidance over the entire course of the project. SMS, TAG, DJD, SK and ET provided clinical expertise and assisted with the interpretation of results.

Funding This work was supported by in part by the National Institutes of Health, National Library of Medicine (grant number 1R01LM011124).

Competing interests None.

Ethics approval The institutional review boards at the Cincinnati Children's Hospital Medical Center, Children's Hospital Colorado, and Children's Hospital of Philadelphia approved this study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The research is funded through an R01 grant, and data use is governed by the protocol.

REFERENCES

1 Hauser WA, Annegers JF, Kurland LT. Incidence of epilepsy and unprovoked seizures in Rochester, Minnesota: 1935–1984. *Epilepsia* 1993;34:453–8.
2 Ivanova JI, Birnbaum HG, Kidolezi Y, *et al*. Direct and indirect costs associated with epileptic partial onset seizures among the privately insured in the United States. *Epilepsia* 2010;51:838–44.
3 Meyer AC, Dua T, Ma J, *et al*. Global disparities in the epilepsy treatment gap: a systematic review. *Bull World Health Organ* 2010;88:260–6.
4 McClelland S III, Guo H, Okuyemi KS. Racial disparities in the surgical management of intractable temporal lobe epilepsy in the United States: a population-based analysis. *Arch Neurol* 2010;67:577.
5 Fountain N, Van Ness P, Swain-Eng R, *et al*. Quality improvement in neurology: AAN epilepsy quality measures: Report of the Quality Measurement and Reporting Subcommittee of the American Academy of Neurology. *Neurology* 2011;76:94–9.

6 Uzuner Ö, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. *J Am Med Inform Assoc* 2009;16:109–15.
7 Fan JW, R Prasad, Yabut RM, *et al*. Part-of-speech tagging for clinical text: wall or bridge between institutions? In: *AMIA Annual Symposium Proceedings*. Vol. 2011. American Medical Informatics Association; 2011:382.
8 Matykiewicz P, Cohen KB, Holland KD, *et al*. In: Earlier Identification of Epilepsy Surgery Candidates Using Natural Language Processing. *ACL*. 2013:1–9.
9 Matykiewicz P, Connolly B, Cohen KB, *et al*. Comparison of corpus linguistics and machine learning techniques in determining differences in clinical notes. Submitted to BMC.
10 Berg AT, Shinnar S, Levy SR, *et al*. How well can epilepsy syndromes be identified at diagnosis? A reassessment 2 years after initial diagnosis. *Epilepsia* 2000;41:1269–75.
11 Aberdeen J, Bayer S, Yeniterzi R, *et al*. The MITRE Identification Scrubber Toolkit: design, training, and assessment. *Int J Med Inform* 2010;79:849–59.
12 Matykiewicz P, Pestian J. Effect of small sample size on text categorization with support vector machines. In: *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. BioNLP '12. Association for Computational Linguistics; 2012:193–201.