Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning

Kevin Bretonnel Cohen[1], Benjamin Glass[2], Hansel M. Greiner[3], Katherine Holland-Bouley[3], Shannon Standridge[3], Ravindra Arya[3], Robert Faist[2], Diego Morita[3], Francesco Mangano[4], Brian Connolly[2], Tracy Glauser[3] and John Pestian[2]

[1]Computational Bioscience Program, University of Colorado School of Medicine, Denver, CO, USA. [2]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, USA. [3]Division of Neurology, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, USA. [4]Division of Pediatric Neurosurgery, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, OH, USA.

ABSTRACT: Objective: We describe the development and evaluation of a system that uses machine learning and natural language processing techniques to identify potential candidates for surgical intervention for drug-resistant pediatric epilepsy. The data are comprised of free-text clinical notes extracted from the electronic health record (EHR). Both known clinical outcomes from the EHR and manual chart annotations provide gold standards for the patient's status. The following hypotheses are then tested: 1) machine learning methods can identify epilepsy surgery candidates as well as physicians do and 2) machine learning methods can identify candidates earlier than physicians do. These hypotheses are tested by systematically evaluating the effects of the data source, amount of training data, class balance, classification algorithm, and feature set on classifier performance. The results support both hypotheses, with F-measures ranging from 0.71 to 0.82. The feature set, classification algorithm, amount of training data, class balance, and gold standard all significantly affected classification performance. It was further observed that classification performance was better than the highest agreement between two annotators, even at one year before documented surgery referral. The results demonstrate that such machine learning methods can contribute to predicting pediatric epilepsy surgery candidates and reducing lag time to surgery referral.

KEYWORDS: epilepsy, epilepsy surgery, neurosurgery, natural language processing, machine learning

CITATION: Cohen et al. Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning. *Biomedical Informatics Insights* 2016:8 11–18 doi: 10.4137/BII.S38308.

TYPE: Original Research

RECEIVED: December 23, 2015. **RESUBMITTED:** March 03, 2016. **ACCEPTED FOR PUBLICATION:** March 03, 2016.

ACADEMIC EDITOR: Phillip Payne, Editorial Board member

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,799 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by the Divisions of Neurology, Neurosurgery and Biomedical Informatics, Cincinnati Children's Hospital Medical Center. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: BG, DM, FM, HG, KBC, KH-B, RA, RF, SS, TG and JP disclose a patent pending for identification of surgery candidates using natural language processing, licensed to Cincinnati Children's Hospital Medical Center (CCHMC). TG and JP disclose a patent issued for processing clinical text with domain specific spreading activation methods, licensed to CCHMC. BC discloses no potential conflicts of interest.

CORRESPONDENCE: john.pestian@cchmc.org

Background and Significance

Epilepsy is characterized by chronic recurrent unprovoked seizures. About 2,000,000 people in the United States and 50 million people worldwide have epilepsy, making it the most common neurological disorder.[1,2]

The current recommendation from the International League Against Epilepsy is that a patient with epilepsy whose seizures have not responded to at least two appropriately chosen and prescribed antiepileptic drugs is considered to have drug-resistant epilepsy, and additional interventions such as epilepsy surgery evaluation should be considered. Epilepsy surgery has become a well-established treatment option for children with drug-resistant epilepsy.[3]

Growing evidence suggests that early surgery has a favorable prognostic implication.[4,5] However, it typically takes several years before a patient with drug-resistant epilepsy is referred for an epilepsy surgery evaluation. Patients who have been referred for epilepsy surgery have sometimes had epilepsy more than 18 years and often more than 10 years after the failure of two antiseizure medications.[6,7] While delays in referral may be most common at primary and secondary care centers, at the institution studied here, the mean time from development of epilepsy to surgery is approximately six years. This indicates that there is work to be done even at a tertiary referral center like this one.

Although the number of epilepsy surgeries in children has steadily increased in the United States in the past two decades, current data suggest that epilepsy surgery is still underutilized. For example, a recent study of trends in epilepsy surgery utilization in children using a national health-care cost and utilization database revealed that less than 35% of expected epilepsy surgeries in children was performed.[8] There are several factors contributing to this underutilization such as lack of clinician awareness of the need and possible outcomes or the family's lack of knowledge regarding this option. A clinical decision support system could help clinicians and families realize that a patient is a potential epilepsy surgery candidate in line with the established American Academy of Neurology standards.[9]

The results in computational techniques suggest that machine learning and natural language processing can be incorporated into decision support systems that help physicians decrease the elapsed time to a surgery referral. Machine learning is a computational technique for developing computer programs that can learn for themselves to make classifications.[10] Natural language processing is a set of computational techniques for using computers to process data that are in the form of language.[11] There are many approaches to natural language processing. In the neurology domain, natural language processing has already been used to capture disease severity from the electronic health record (EHR) in multiple sclerosis[12]; several decision support tools are available in neurological disease, including Simulconsult (www.simulconsult.com). However, these methods have as yet not been integrated into the clinical workflow.

This work is innovative from both the clinical neurology perspective and the informatics perspective. From a clinical perspective, it could form the basis of the first clinical decision support system for epilepsy surgery treatment. From an informatics perspective, the analysis is unusual in that it systematically explores the impact of a number of factors on a text classification task, ie, data source, data size, data balance, classification algorithm, and feature set. It is also unusual in that it reports the measures of dispersion of the figures of merit. Of course, many of these factors have been examined before, but not typically all of them and not typically in the context of a clinically relevant task.

Materials and Methods

EHR data. This study was approved by the Cincinnati Children's Hospital Medical Center's (CCHMC's) Institutional Review Board. We collected a data set of free-text clinical notes by querying the electronic medical record at CCHMC, a large academic pediatric neurology practice.

We first identified all current and past patients who were assigned International Classification of Disease version 9 (ICD-9) codes for epilepsy or convulsions.[†] Only patients who had seen a physician or practitioner in the Division of Neurology for at least one year and had at least four progress notes since 2009 were included. A total of 6,343 patients fulfilled these criteria.

From this group, positive and negative examples were identified for training an epilepsy surgery candidate classifier.

- Positive instances were patients who underwent resective epilepsy surgery, defined as either lobectomy, corticectomy, or hemispherectomy. These were identified using procedure codes corresponding to craniotomies for resection[‡] ($N = 100$).
- Negative instances were nonsurgical patients identified as seizure free for at least 12 months before their latest visit, using the last seizure date found in structured and unstructured EHR data ($N = 423$).

Manually annotated subset. To assess human performance on this task, a group of pediatric epileptologists manually annotated a subset of the data with the classes that the system is intended to assign. The labeling was done at the patient level and at the level of each clinic visit note, corresponding to that patient.

A sample of 62 patients with unknown outcomes and their clinic visit notes were chosen for manual annotation. Possible annotations were "surgery candidate", "non-candidate", and "undetermined". Manual labeling was done by three clinicians, with a fourth (senior) neurologist resolving cases where there was no agreement between any of the three raters. The product of this annotation is what is used to calculate the inter-annotator agreement numbers given in the "The context for interpretation: baseline and upper bound" section.

Natural language processing and machine-learning algorithms. The experimental design and subsequent analysis were driven by the following general principles:

1. Since interpreting system performance depends on the task, it is necessary to choose a baseline algorithm that defines the lower bound for performance. The most valid baseline is a baseline internal to this work. That is, we aim for an apples-to-apples comparison of the final system to a baseline that we built, keeping as much of the system constant as possible, other than the aspects of the system whose impact is being evaluated by any given experiment.
2. The most reasonable upper bound against which to evaluate system outputs is the inter-annotator agreement.
3. Evaluation of Hypothesis 1 – that the system can identify surgical candidates at the same level of performance as the physicians – is best done with a single value for the figure of merit.
4. Evaluation of Hypothesis 2 – that the system can identify surgical candidates earlier than the physicians – is done by examining the time course of the figure of merit.

The analysis examines the effects of the following on system performance:

1. Size of training data,
2. Feature set,
3. Classification algorithm,
4. Balance of positive and negative classes, and

[†] 345.*, 780.3*, 779.0

[‡] 61510, 61531, 61533–61540, 61542–61543, 61566–61567

5. Manually labeled (annotated) versus weakly labeled (EHR) data.

Assuming three differently sized data sets, five feature sets, and two classification algorithms (holding all the parameters of the algorithm constant at the default values) would give 30 sets of results for a reasonable test of either of the hypotheses. This does not begin to exhaust the parameter space, as we have shown elsewhere,[13] even with a number of simplifying assumptions, just exploring the support vector machine (SVM)-light (http://svmlight.joachims.org) functionality alone would require 92,980,917,360 different sets of parameter settings. If we could execute one run per second, it would take a bit under 3,000 years to run the experiment (we are aware that there are algorithmic approaches to making the process more efficient but do not know of any such approaches that have been attempted on the problem of this scale), and it is unlikely that the results would be illuminating in any scientific sense. (It is not impossible that they would be: in Ref 14 we showed that exhaustive exploration of a large parameter space reveals an interesting pattern in the relationship between precision and recall on the gene mention task, on one hand, and performance on the gene normalization on the other.) We present here a rational subset of factors that seemed likely to have a computationally, statistically, or linguistically motivated effect on performance, while striving to maintain the apples-to-apples comparisons that we identified above as desiderata of the work. Specifically, we compare the following.

Size of training data. In one experiment, we varied the amount of training data, from 40 patients to 200 patients. We held all other factors constant: balanced data, feature set (unigrams + bigrams + drugs), classifier type (SVM), and data source (distant supervision).

Feature set. In another experiment, we varied the feature set. Specifically, we compared the following:

- Unigrams;
- Unigrams + drug name normalization;
- Bigrams;
- Bigrams + drug name normalization;
- Unigrams + bigrams; and
- Unigrams + bigrams + drug name normalization.

These were chosen on the basis of the following rationale. Despite common assumptions to the contrary, it is not an a priori given that bigrams will yield an improvement over unigrams.[15,16] For example, when adding bigrams to unigrams alone, either no improvement or worsening of performance was observed.[17–19] Bekhuis and Demner-Fushman[19] also point out that although bigrams often do improve performance, unigrams can decrease the computational load. In these experiments, unigrams were used because they are a nontrivial baseline for document classification.[19–26] Bigrams were added because the combination of unigrams and bigrams is a common feature set in document classification. Drug name normalization was added because medication names are readily extractable and have clear clinical relevance. We mapped all known generic or trade names of 42 different antiepileptic drugs to codes, representing the medications. For example, *lacosamide* and *Vimpat* were both replaced with *_DRUG_ LCM*. All other factors such as balanced data, the size of the training set (200 patients), classifier type (SVM), and data source (distant supervision) were held constant.

Classification algorithm. To examine the effect of the choice of the classification algorithm, we compared a Naive Bayes classifier with an SVM. In the machine-learning literature, in general, this is not an uncommon comparison.[19,27–29] We used a Naive Bayes classifier because it is a nontrivial baseline for document classification tasks.[11,23,30–32] We compared it with a SVM because of the long history of strong performance of SVMs in document classification, the public availability of excellent implementations of SVM classifiers (which we hoped would enable reproducibility), and our long experience with this type of classifier. We used the MALLET implementation[33] of Naive Bayes and the scikit-learn[34] implementation of a SVM. All other factors such as balanced data, the size of the training set (200 patients), feature set (unigrams + bigrams + drugs), and data source (distant supervision) were held constant.

Balance of positive and negative classes. We examined the effect of balance in the training and testing data by varying the ratio of positive to negative instances (that is, surgery candidates versus noncandidates) from 1:1 to 1:4. The overall size of the data set increased from 200 to 500. The feature set (unigrams + bigrams + drugs), the classification algorithm (SVM), and data source (distant supervision) were held constant.

Manually labeled versus weakly labeled data. Finally, we evaluated the effect of manually labeled (annotated) versus weakly labeled data. A set of four board-certified pediatric neurologists manually annotated data from 62 patients. Patients and notes were classified as being (in the case of the patients) or indicating (in the case of their notes) intractable/intractability, epilepsy type, and surgical candidates/candidacy. It is a classic example of what has been called light annotation – that is, annotation requiring domain expertise and yielding domain labels, as opposed to linguistic labels (such as part of speech or syntactic structure).[35,36] We refer these data as manually labeled or manually annotated data.

We compared classifier performance with these data to classifier performance with data pulled from the hospital's EHR system using the criteria described elsewhere in this article. These kinds of data are typically called weakly labeled data.[20,37] It has inspired the approach to data use known as distant supervision that has recently achieved wide currency in the broader language processing world.[38]

Results

The context for interpretation: baseline and upper bound. We report classification results primarily in the form of *F*-measure, defined as the harmonic mean of precision and recall.

In considering the results of various experiments, it is important to have a clear picture of the upper and lower bounds of likely performance.

The baseline is derived from a system using unigrams (unordered single words) as features, a Naive Bayes classifier, and training and testing on 100 positive (surgical) and 100 negative (nonsurgical) patients with 10-fold cross-validation. This is a nontrivial baseline, and often, it is not an easy one to beat. The resulting baseline is $F = 0.74 \pm 0.04$; any system that does not perform better than this is failing in some way.

The upper bound for the system is the level of agreement of human annotators with each other. For the purpose of measuring agreement, the annotations were collapsed to a binary scheme: surgery candidates versus noncandidates or undetermined candidacy. Of the 62 manually annotated patients, 19 were annotated as candidates and 43 as noncandidates or undetermined. For the two annotators with the highest agreement, Cohen's kappa was 0.632 and the *F*-measure (considering either annotator as the gold standard) was 0.71 ± 0.10. The relatively poor agreement can be attributed to varying interpretations of incomplete EHR data in many of the patient records. Any system approaching, and certainly any system exceeding, the agreement *F*-measure of 0.71 can be considered to be performing as well as is possible.

Patient-level classification. Table 1 lists the results for the baseline system at the most basic analytical level – that is, the classification of patients (n = 200). The baseline system performs comparably to the aggregate of neurologists. The aggregate has an inter-annotator agreement (which, as described above, is a likely ceiling on possible performance) *F*-measure of 0.71 ± 0.10, while the baseline system performs at 0.74 ± 0.04, supporting Hypothesis 1.

Temporal variation in classification. Figure 1 shows classification results through 11 successive clinic visits, as the system takes more and more data for each patient into account. Figure 2 shows similar data, but with results reported over nine time periods rather than numbered office visits. These results test the hypothesis that the system can identify surgical candidates earlier than physicians. In both cases, the system performs comparably to the aggregate of physicians from the first time period and improves over time, supporting Hypothesis 2.

Effect of feature set. Analyzing the effect of different feature sets gives us some idea of how much potential there might be for a feature engineering-based approach to improving the system without overfitting.

We evaluated various combinations of surface linguistic features and conceptual features. The surface linguistic features were unigrams and bigrams. As a conceptual feature, we looked at the recognition of drug names.

Table 1. Patient level classification of surgery candidacy, baseline system. ± values are standard deviations. The baseline system performs comparably to the aggregate of neurologists.

		NLP		
		SURGERY CANDIDATE	NON-CANDIDATE	TOTAL
GOLD STANDARD	SURGERY CANDIDATE	71	29	100
	NON-CANDIDATE	21	79	100
	TOTAL	92	108	200

Notes: Precision: 0.77 ± 0.03. Recall: 0.71 ± 0.03. F-measure: 0.74 ± 0.04.

The results for various combinations of these features are listed in Table 2. In these experiments, we kept all other aspects of the processing constant as follows: balanced data (100 positive and 100 negatives), classification algorithm (SVM), and data source (distant supervision). Overall, unigrams only are a strong baseline, performing better than both the inter-annotator agreement and the Naive Bayes baseline system. Any additional features improve on the unigram-only performance. In particular, drug recognition improves on all combinations of surface linguistic features, suggesting the addition of additional conceptual classes as an avenue for improving performance further.

Classification algorithm. We evaluated two different classification algorithms to get some insight into the likelihood of future benefit from trying a wider range of classification algorithms. We compared Naive Bayes and a SVM. We kept all other aspects of the processing constant as follows: balanced data (100 positive and 100 negatives), feature set (unigrams, bigrams, and drugs), and data source (distant supervision). Table 3 shows the results. The SVM produced considerably higher performance than the Naive Bayes classifier.

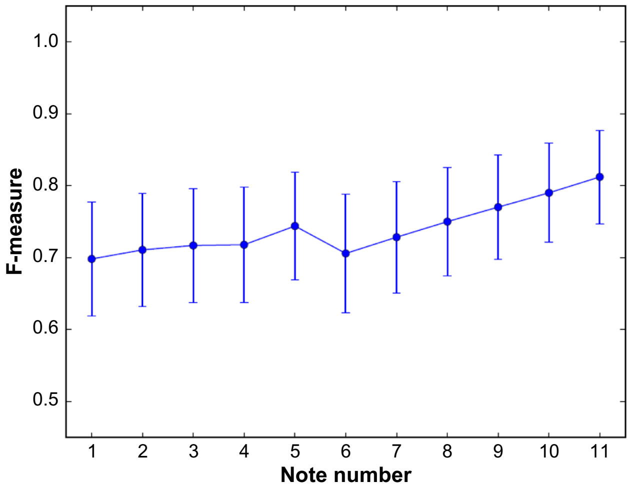

Figure 1. Classification of surgery candidacy at 11 clinic visits by the baseline system. Error bars are 95% confidence intervals. The mean performance of the baseline system is comparable to the aggregate of neurologists at the first visit and improves over time.

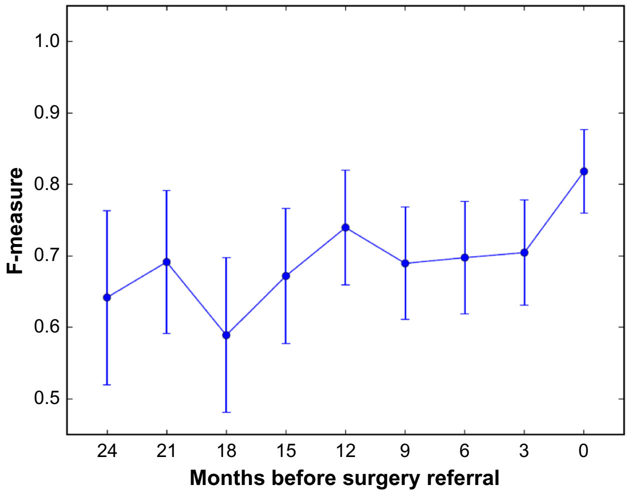

Figure 2. Classification of surgery candidacy at nine time periods by the baseline system. Error bars are 95% confidence intervals. The mean performance of the baseline system is comparable to the aggregate of neurologists at the first time period and improves over time.

Effect of size of training data. To assess how much performance gain we might see from acquiring additional data as compared with the cost of getting that data, we examined the learning curve, as increasing amounts of data are used to train the classifier. We held all other factors constant: balanced data, feature set (unigrams + bigrams + drugs), classifier type (SVM), and data source (distant supervision).

Examining the results in Figure 3, we see that the performance at a training set size of 200 (the maximum that we were able to do with our data set) has already been reached at a training set size of 120. Thus, it seems likely that additional data might improve the results, but that quite a bit of data would be needed to see an improvement.

Effect of balance. The system was quite sensitive to the balance of the data. With a 1:1 ratio of positive:negative instances (that is, candidates to noncandidates) and the configuration described in the "Material and methods" section, the system achieved an *F*-measure of 0.82 ± 0.03. With a 1:4 ratio, the *F*-measure dropped to 0.70 ± 0.04. Table 4 shows a rapid drop-off in performance, as the balance changes from 1:1 to 1:4. This is especially notable since as the balance changes from 1:1 to 1:4, the size of the training set more than doubles.

Effect of data source. We evaluated the effect of varying the data source. We compared the models trained on data from the two sources, holding every other factor constant: unbalanced data (19 positive, 43 negative), feature set (unigrams + drugs), and classifier (SVM). Table 5 shows the results. With these parameters, the distant supervision approach yielded better results than did the manually annotated data.

Table 2. *F*-measure with various feature sets.

FEATURE SET	*F*-MEASURE
Unigrams	0.77 ± 0.03
Unigrams + drugs	0.81 ± 0.03
Bigrams	0.80 ± 0.03
Bigrams + drugs	0.80 ± 0.03
Unigrams + bigrams	0.80 ± 0.03
Unigrams + bigrams + drugs	0.82 ± 0.03

Note: All other factors such as balanced data, data set size, classification by SVM, and distant supervision for the data source are held constant.

Table 3. Effect of classifier, holding all other factors constant.

CLASSIFIER	*F*-MEASURE
Naive Bayes	0.77 ± 0.03
Support vector machine	0.82 ± 0.03

Discussion

These experiments evaluate two hypotheses: that an automated system can identify surgical candidates as well as board-certified neurologists and that the automated system can identify surgical candidates earlier than neurologists. The data shown in Table 1 suggest that the first hypothesis is supported. The data shown in Figures 1 and 2 suggest that the second hypothesis is also supported.

We sought to better understand what factors affect the performance of the system, and in what way. Our experiments varied specific things that could be expected to affect performance, holding all other aspects of the system constant. The results of these experiments were consistent with the following analyses:

- Feature engineering pays off for these data. The simplest set of features yielded an *F*-measure of 0.77, and the broadest set of features yielded an *F*-measure of 0.82.

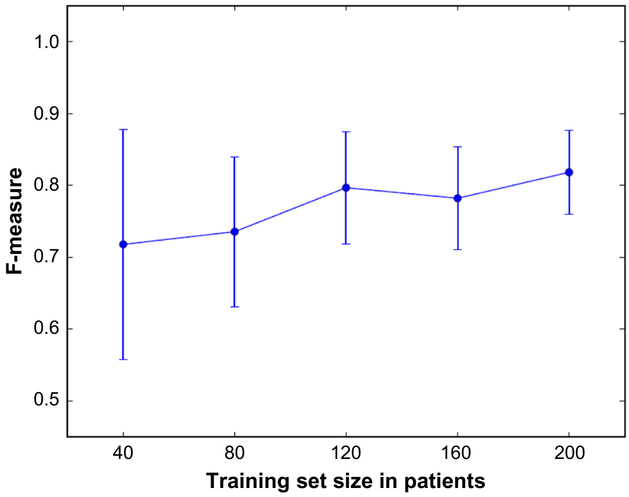

Figure 3. Effect of size of training data. All other factors such as balance, feature set, classifier type, and data source are held constant. Error bars are 95% confidence intervals.

Table 4. Effect of data balance, holding all other factors constant.

POSITIVE INSTANCES	NEGATIVE INSTANCES	*F*-MEASURE
100	100	0.82 ± 0.03
100	200	0.80 ± 0.03
100	300	0.74 ± 0.04
100	400	0.70 ± 0.04

These figures reflect the results of only a very modest amount of feature engineering, and it is likely that more such effort would yield increased performance.

- Classification algorithm does matter. The baseline Naive Bayes system yielded an *F*-measure of 0.77, while the SVM produced an *F*-measure of 0.82.
- Training data size matters. A fivefold increase in the number of patients in the data set (from 40 to 200) yielded an increase of 10 points of *F*-measure (from 0.718 to 0.818). This is a large increase, but perhaps not proportional to the increase in the size of the data set, although certainly consistent with other findings in the literature, as discussed below. This suggests that large amounts of additional data could improve performance further.
- Balance does matter. The performance was considerably higher with balanced data than with unbalanced data – 0.70 *F*-measure to 0.82 *F*-measure.
- Performance is affected by the source of the training data. The manual annotation may not be worth the expense with such a small data set, as is the case here. It is not unreasonable to consider the impact of larger corpora.

One way in which this work differs from the broader body of literature in natural language processing is that we have reported the statistical dispersion of the figures of merit. These figures of dispersion demonstrate the importance of reporting statistical variability in natural language processing and machine learning, and in fact, they suggest some caution in interpreting the results. For example, looking at the error bars in Figures 1–3, it is apparent that for some measurement points, there is a substantial difference between the mean performance of the system and its extreme points and that changes (and particularly improvements) in performance that seem clear from the trends in mean performance do not seem so evident when considering the variability in performance. This point is not typically considered in the related literature; the results described here suggest that it should be.

Table 5. Effect of data source, holding all other factors constant.

DATA SOURCE	*F*-MEASURE
Distant supervision	0.74 ± 0.08
Manually annotated	0.70 ± 0.08

Comparison to the literature. Machine learning has had some applications in neurology. For example, Memarian et al.[39] used machine learning for epilepsy surgery outcome prediction, Dian et al.[40] used machine learning to find brain regions of interest for surgery, and Yang et al.[41] used machine learning to characterize lateralization. Jette et al.[42] and Roberts et al.[43] report an online tool for evaluating appropriateness for epilepsy surgery evaluation. The system reported here differs in that it can identify patients proactively and a tool can be developed to automatically notify the provider. This can be integrated into the normal workflow of patient care.

Matykiewicz et al provided a proof of concept for the idea that early prediction of pediatric epilepsy surgery candidates might be possible.[44] The work reported here is a thorough exploration of the methods that could potentially be applied to that task. While Matykiewicz et al.[44] reported on a single classifier, this work explores the factors that affect such models and provides insights into how such models can be improved – and made practical. Although we have not made scaling issues a focus of this article, we note a relevant comparison in Ref. 44 with respect to this topic. The system described in that study achieved similar results, but with four days of training and testing time. In contrast, one run of the system described here takes less than an hour. Used in the clinic on a patient-by-patient basis, this system could return results before the bloodwork is back from the laboratory.

Perhaps the biggest surprise of the results reported here was the amount of data required to improve the performance above the baseline. It has been known since the publication of Ref. 45 that the amount of data available to a training algorithm has a strong effect on classifier performance in natural language processing. Banko and Brill[45] set out to evaluate the effect of varying training data set size on a classifier. The goal was to determine at what point classifier performance asymptotes, as the amount of training data is increased. Using the training set data of gradually increasing sizes, up to 1,000 times the size of the previously largest training set, they found that performance might never asymptote. All of the five classifiers that they tried showed improvements, sometimes in surprising ways – for example, the worst system with a training set of 1 million words is the second best with 1 billion words. The authors concluded that "it may make sense for the field to concentrate considerably more effort into enlarging our training corpora and addressing scalability issues, rather than continuing to explore different learning methods".[45] Today, it is a commonplace for machine-learning papers to include a discussion of the "learning curve," or the way that performance changes as the training set size is increased, eg, for recent work in the biomedical domain.[46–48] Getting enough data to increase performance beyond what we have seen is likely to require multicenter studies.

Abundant research indicates that the set of features used for a particular classification task is crucial to the results.[49–54] Evaluation of feature sets in natural language processing can be a challenge, to do in a coherent and comprehensible way. The number of possible feature sets that could be evaluated is equal to 2 raised to the number of types of features. The standard approach for any number of types of features larger than 3 or so is to pick a rational subset of the possible feature sets, with new features being added to some baseline set. We took that approach here and found that the feature set did make a large difference in performance, with *F*-measure varying from 0.77 to 0.82 depending on which features were used.

In a study of classification of systematic reviews,[19] it was found that no classifier achieved sufficient recall without the optimization of classification parameters. That work involved extensive tuning. In particular, for the SVM, the effects of kernel type, presence or absence of Gaussian mutation, gamma value, epsilon value, population size, and *C* were optimized. For the Naïve Bayes classifier, smoothing values and normalized class weights were optimized. In contrast, the work reported in this article involved optimizing only kernel type, *C*, and gamma. Based on Bekhuis and Demner-Fushman's[19] experience, optimization seems likely to improve the performance of the baseline system reported here.

Feature sets that take advantage of biomedical document structure can yield large performance increases.[19,55–57] Although these features can be very genre specific and thus difficult to generalize, there is some reason to think that systems for automatically determining document structure are within reach,[58] making this a plausible avenue for future research.

In addition to surface linguistic features (such as unigrams and bigrams), conceptual features have often been found useful in classification tasks in natural language processing. By conceptual features, we mean features that reflect the presence of some abstract class in a document, such as weights,[59] drugs (the present work), smoking status,[60] and genes.[61] The effect of inclusion of drugs in the feature set experiments suggests that additional conceptual features will produce more gains in performance.

Conclusion

The results of the structured experiments examining the effects of training data size, balance, feature sets, and classification algorithm suggest a path toward improved performance on the classification task for these kinds of clinical data. In particular, the feature set is a strong contributor to performance, and adding additional features – particularly additional conceptual features – appears to be a strategy with a high likelihood of success.

In the context of the broader literature on classification in natural language processing, in general, and in the biomedical domain, in particular, the work reported here is motivated (in addition to its clinical significance) by the observation that the results of classification experiments need to be validated on other tasks.[19,29] This study has defined a strong baseline approach to the task of predicting candidacy for surgical intervention in pediatric epilepsy patients. Additionally, it has established the effects of a number of factors on the performance of systems that carry out this task. It suggests some avenues for improving performance and holds promise for the development of tools to support physicians in the work of pediatric neurology. By applying this system to all patients with epilepsy, a clinician could improve his/her recognition of intractability and decrease the lag time for a referral for this potentially curative intervention. Additionally, this system could be used in a predictive fashion and help identify patients who are at risk of developing intractability, which may in the future change our current management approach.

Author Contributions

Conceived and designed the experiments: KBC, BG, BC, JP. Developed software and analyzed the data: KBC, BG, RF. Annotated data: HMG, KHB, RA, TG. Wrote the first draft of the manuscript: KBC, BG, HMG, DM, JP. Made critical revisions: KHB, SS, RA, FM, BC, TG, JP. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Chang BS, Lowenstein DH. Epilepsy. *N Engl J Med*. 2003;349:1257–66.
2. Organization WH. *Epilepsy: Fact Sheet #999*. Available at: http://www.who.int/mediacentre/factsheets/fs999/en/. Accessed November 04, 2015.
3. Hemb M, Velasco T, Parnes M, et al. Improved outcomes in pediatric epilepsy surgery the UCLA experience, 1986–2008. *Neurology*. 2010;74(22):1768–75.
4. Janszky J, Janszky I, Schulz R, et al. Temporal lobe epilepsy with hippocampal sclerosis: predictors for long-term surgical outcome. *Brain*. 2005;128(2):395–404.
5. Simasathien T, Vadera S, Najm I, Gupta A, Bingaman W, Jehi L. Improved outcomes with earlier surgery for intractable frontal lobe epilepsy. *Ann Neurol*. 2013;73(5):646–54.
6. Berg A, Langfitt J, Shinnar S, et al. How long does it take for partial epilepsy to become intractable? *Neurology*. 2003;60(2):186–90.
7. Haneef Z, Stern J, Dewar S, Engel J. Referral pattern for epilepsy surgery after evidence-based recommendations: a retrospective study. *Neurology*. 2010;75(8):699–704.
8. Pestana Knight EM, Schiltz NK, Bakaki PM, Koroukian SM, Lhatoo SD, Kaiboriboon K. Increasing utilization of pediatric epilepsy surgery in the United States between 1997 and 2009. *Epilepsia*. 2015;56(3):375–81.
9. Fountain NB, Van Ness PC, Swain-Eng R, Tonn S, Bever CT Jr; American Academy of Neurology Epilepsy Measure Development Panel and the American Medical Association-Convened Physician Consortium for Performance Improvement Independent Measure Development Process. 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(1):94–9.
10. Mitchell TM. *The Discipline of Machine Learning*. Vol 9. Carnegie Mellon University, School of Computer Science, Machine Learning Department. Pittsburgh, PA, USA; 2006.
11. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Prentice Hall. Upper Saddle River, NJ, USA; 2008.
12. Xia Z, Secor E, Chibnik LB, et al. Modeling disease severity in multiple sclerosis using electronic health records. *PLoS One*. 2013;8(11):e78927.
13. Cohen KB, Hunter LE, Palmer M. Assessment of software testing and quality assurance in natural language processing applications and a linguistically inspired approach to improving it. Alessandro Moschitti, Barbara Plank, Editors. In: *Trustworthy Eternal Systems via Evolving Software, Data and Knowledge*. Springer, Berlin DE; 2013:77–90.
14. Baumgartner WA Jr, Cohen KB, Hunter L. An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. *J Biomed Discov Collab*. 2008;3:1.
15. Tan CM, Wang YF, Lee CD. The use of bigrams to enhance text categorization. *Inf Process Manage*. 2002;38(4):529–46.
16. Bekkerman R, Allan J. *Using Bigrams in Text Categorization*. Vol 1003. Amherst: Department of Computer Science, University of Massachusetts; 2004:1–2.

17. Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*. Association for Computational Linguistics. Philadelphia, PA, USA; 2002:79–86.
18. Minard AL, Ligozat AL, Abacha AB, et al. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *J Am Med Inform Assoc*. 2011;18(5):588–93.
19. Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artif Intell Med*. 2012;55(3):197–207.
20. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In: *Intelligent Systems for Molecular Biology*. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. Heidelberg, DE. AAAI Press; 1999:77–86.
21. Manning C, Schuetze H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA; 1999.
22. Tillmann C. A unigram orientation model for statistical machine translation. In: *Proceedings of HLT-NAACL 2004: Short Papers*. Association for Computational Linguistics,Boston, MA, USA; 2004:101–4.
23. Jackson P, Moulinier I. Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization. 2nd ed. John Benjamins Publishing Company, Amsterdam, Netherlands; 2002.
24. Li Q, Zhai H, Deleger L, et al. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. *J Am Med Inform Assoc*. 2013;20(5):915–21.
25. Marafino BJ, Davies JM, Bardach NS, Dean ML, Dudley RA. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *J Am Med Inform Assoc*. 2014;21(5):871–5.
26. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. *J Am Med Inform Assoc*. 2015;22(1):155–65.
27. Han H, Giles L, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. In: Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries. IEEE,Tucson, AZ, USA; 2004:296–305.
28. Cesa-Bianchi N, Gentile C, Zaniboni L. Hierarchical classification: combining Bayes with SVM. In: *Proceedings of the 23rd International Conference on Machine learning*. ACM, Pittsburgh, PA, USA; 2006:177–84.
29. Colas F, Brazdil P. Comparison of SVM and some older classification algorithms in text classification tasks. In: *Artificial Intelligence in Theory and Practice*. Springer, New York, NY, USA; 2006:169–78.
30. Kim SB, Rim HC, Yook D, Lim HS. Effective methods for improving naive Bayes text classifiers. In: *PRICAI 2002: Trends in Artificial Intelligence*. Springer, New York, NY, USA; 2002:414–23.
31. Rennie JD, Shih L, Teevan J, et al. Tackling the poor assumptions of naive Bayes text classifiers. In: *ICML*. Vol 3. Omnipress, Madison WI, Washington, DC; 2003:616–23.
32. Kim SB, Han KS, Rim HC, Myaeng SH. Some effective techniques for naive Bayes text classification. *EEE Trans Knowl Data Eng*. 2006;18(11):1457–66.
33. McCallum AK. {MALLET: A Machine Learning for Language Toolkit}; 2002. http://mallet.cs.umass.edu
34. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
35. Stubbs A. A Methodology for Using Professional Knowledge in Corpus Annotation. Brandeis University, Waltham, MA, USA; 2013.
36. Pustejovsky J, Stubbs A. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Sebastopol, CA, USA; 2012.
37. Morgan AA, Hirschman L, Colosimo M, Yeh AS, Colombe JB. Gene name identification and normalization using a model organism database. *J Biomed Inform*. 2004;37(6):396–410.
38. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Vol 2. Association for Computational Linguistics, Singapore, Singapore; 2009:1003–11.
39. Memarian N, Kim S, Dewar S, Engel J, Staba RJ. Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. *Comput Biol Med*. Elsevier, Philadelphia, PA, USA; 2015;64:67–78.
40. Dian JA, Colic S, Chinvarun Y, Carlen PL, Bardakjian BL. Identification of brain regions of interest for epilepsy surgery planning using support vector machines. In: *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE; 2015:6590–3.
41. Yang Z, Choupan J, Reutens D, Hocking J. Lateralization of temporal lobe epilepsy based on resting-state functional magnetic resonance imaging and machine learning. *Front Neurol*. 2015;6:184.
42. Jette N, Quan H, Tellez-Zenteno JF, et al. Development of an online tool to determine appropriateness for an epilepsy surgery evaluation. *Neurology*. 2012;79(11):1084–93.
43. Roberts JI, Hrazdil C, Wiebe S, et al. Feasibility of using an online tool to assess appropriateness for an epilepsy surgery evaluation. *Neurology*. 2014;83(10):913–9.
44. Matykiewicz P, Cohen KB, Holland KD, et al. Earlier identification of epilepsy surgery candidates using natural language processing. *ACL*. 2013;2013:1.
45. Banko M, Brill E. Scaling to very large corpora for natural language disambiguation. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Toulouse, France; 2001:26–33.
46. Chen Y, Cao H, Mei Q, Zheng K, Xu H. Applying active learning to supervised word sense disambiguation in MEDLINE. *J Am Med Inform Assoc*. 2013;20(5):1001–6.
47. Chen Y, Carroll RJ, Hinz ERM, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *J Am Med Inform Assoc*. 2013;20(e2):e253–9.
48. Suominen H, Johnson M, Zhou L, et al. Capturing patient information at nursing shift changes: methodological evaluation of speech recognition and information extraction. *J Am Med Inform Assoc*. 2015;22(e1):e48–66.
49. Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. *J Am Med Inform Assoc*. 2004;11(4):320–31.
50. Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. *J Am Med Inform Assoc*. 2012;19(5):817–23.
51. Zheng J, Chapman WW, Miller TA, Lin C, Crowley RS, Savova GK. A system for coreference resolution for the clinical narrative. *J Am Med Inform Assoc*. 2012;19(4):660–7.
52. Grouin C, Grabar N, Hamon T, Rosset S, Tannier X, Zweigenbaum P. Eventual situations for timeline extraction from clinical reports. *J Am Med Inform Assoc*. 2013;20(5):820–7.
53. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc*. 2015;22(3):671–81.
54. Pradhan S, Elhadad N, South BR, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc*. 2015;22(1):143–54.
55. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical text. *J Am Med Inform Assoc*. 2007;14(3):253–63.
56. Deléger L, Grouin C, Zweigenbaum P. Extracting medical information from narrative patient records: the case of medication-related information. *J Am Med Inform Assoc*. 2010;17(5):555–8.
57. Clark C, Aberdeen J, Coarr M, et al. MITRE system for clinical assertion status classification. *J Am Med Inform Assoc*. 2011;18(5):563–7.
58. Demner-Fushman D, Abhyankar S, Jimeno-Yepes A, et al. A knowledge-based approach to medical records retrieval. In: *TREC,* Gaithersburg, MD, USA; 2011.
59. Patel CO, Cimino JJ. Using semantic and structural properties of the unified medical language system to discover potential terminological relationships. *J Am Med Inform Assoc*. 2009;16(3):346–53.
60. Abhyankar S, Leishear K, Callaghan FM, Demner-Fushman D, McDonald CJ. Lower short-and long-term mortality associated with overweight and obesity in a large cohort study of adult intensive care unit patients. *Crit Care*. 2012;16(6):R235.
61. Baumgartner WA Jr, Lu Z, Johnson HL, et al. Concept recognition for extracting protein interaction relations from biomedical text. *Genome Biol*. 2008;9(suppl 2):S9.