Received: 20 July 2019 | Revised: 5 November 2019 | Accepted: 5 November 2019

DOI: 10.1111/epi.16398

Check for updates

Epilepsia®

FULL-LENGTH ORIGINAL RESEARCH

Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery

Benjamin D. Wissel[1] | **Hansel M. Greiner**[2,3] | **Tracy A. Glauser**[2,3] | **Katherine D. Holland-Bouley**[2,3] | **Francesco T. Mangano**[2,4] | **Daniel Santel**[1] | **Robert Faist**[1] | **Nanhua Zhang**[2,5] | **John P. Pestian**[1,2] | **Rhonda D. Szczesniak**[2,5] | **Judith W. Dexheimer**[1,2,6]

[1]Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

[2]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio

[3]Division of Neurology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

[4]Division of Neurosurgery, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

[5]Division of Biostatistics & Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

[6]Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio

Correspondence
Judith Dexheimer, Division of Emergency Medicine, Cincinnati Children's Hospital Medical Center, MLC 2008, 3333 Burnet Avenue, Cincinnati, OH 45229-3039.
Email: Judith.Dexheimer@cchmc.org

Funding information
National Heart, Lung, and Blood Institute, Grant/Award Number: K25 HL12595; Agency for Healthcare Research and Quality, Grant/Award Number: R21 HS024977

Abstract

Objective: Delay to resective epilepsy surgery results in avoidable disease burden and increased risk of mortality. The objective was to prospectively validate a natural language processing (NLP) application that uses provider notes to assign epilepsy surgery candidacy scores.

Methods: The application was trained on notes from (1) patients with a diagnosis of epilepsy and a history of resective epilepsy surgery and (2) patients who were seizure-free without surgery. The testing set included all patients with unknown surgical candidacy status and an upcoming neurology visit. Training and testing sets were updated weekly for 1 year. One- to three-word phrases contained in patients' notes were used as features. Patients prospectively identified by the application as candidates for surgery were manually reviewed by two epileptologists. Performance metrics were defined by comparing NLP-derived surgical candidacy scores with surgical candidacy status from expert chart review.

Results: The training set was updated weekly and included notes from a mean of 519 ± 67 patients. The area under the receiver operating characteristic curve (AUC) from 10-fold cross-validation was 0.90 ± 0.04 (range = 0.83-0.96) and improved by 0.002 per week ($P < .001$) as new patients were added to the training set. Of the 6395 patients who visited the neurology clinic, 4211 (67%) were evaluated by the model. The prospective AUC on this test set was 0.79 (95% confidence interval [CI] = 0.62-0.96). Using the optimal surgical candidacy score threshold, sensitivity was 0.80 (95% CI = 0.29-0.99), specificity was 0.77 (95% CI = 0.64-0.88), positive predictive value was 0.25 (95% CI = 0.07-0.52), and negative predictive value was 0.98 (95% CI = 0.87-1.00). The number needed to screen was 5.6.

Significance: An electronic health record–integrated NLP application can accurately assign surgical candidacy scores to patients in a clinical setting.

KEYWORDS

clinical decision support, epilepsy surgery, machine learning, natural language processing

© 2019 International League Against Epilepsy

1 | INTRODUCTION

Approximately 30% of patients with epilepsy are drug-resistant. A select portion of patients with drug-resistant epilepsy are candidates for resective surgery.[1–3] Patients who undergo resective surgery have a 58%-78% chance of seizure freedom, compared to only 7% with appropriate medical therapy.[4–8] Children have an improved quality of life after surgery, and earlier surgery further improves cognitive and seizure outcomes.[6,9,10] Surgery morbidity is modest[11]; however, surgical care remains underutilized despite referral guidelines.[12–15] Epilepsy disease duration before surgery still exceeds 6 years in pediatric patients and 20 years in adults.[11] Half of neurologists are unsure when to refer patients with epilepsy for surgical consultation.[16] Reliable identification of candidates for resective epilepsy surgery could facilitate earlier referral for surgical treatment.

Neurologists' opinions of which patients should be referred for surgery are varied.[17] In response, an online clinical decision support tool was created, the Canadian Appropriateness of Epilepsy Surgery (CASES). CASES evaluates the appropriateness of surgical referrals and has a sensitivity of 95% for identifying guideline-eligible patients.[14,18,19] CASES requires providers to manually enter data. Automating this process to eliminate the need for manual entry of data into a decision support tool would enable surgical candidacy to be evaluated in more patients with less interruption of physician workflow.[20,21]

Natural language processing (NLP) is a branch of artificial intelligence that analyzes unstructured free text.[22] We developed an NLP application to identify patients who are potentially eligible for presurgical evaluations for resective epilepsy surgery.[23,24] The application identified surgical consultation eligible patients up to 2 years prior to their referral.[23,24] Our model used provider notes and achieved equal classification accuracies to those of epileptologists in a direct comparison.[23] It generated surgical candidacy scores that were not biased by patients' gender or race.[25] We integrated our model with a pediatric hospital's electronic health record (EHR) and assigned surgical candidacy scores to epilepsy patients who were scheduled to visit the neurology clinic. The objective of this study was to prospectively validate the performance of the NLP application in a clinical setting.

Key Points

- A retrospectively developed natural language processing application can prospectively identify potential candidates for resective epilepsy surgery
- The application learned to assign weights to key words and phrases without needing to incorporate a priori domain knowledge
- The application's performance increased throughout the 1-year study period as new patients were added to the training set
- The approach is sustainable over time
- The application may be implemented to generate electronic alerts for potential surgical candidates

2 | MATERIALS AND METHODS

2.1 | Ethics and waiver of informed consent

The institutional review board at Cincinnati Children's Hospital Medical Center (CCHMC) approved the study with a waiver of informed consent. This article complies with the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) reporting guidelines.[26] The notes used in this study cannot be made freely available to the public, because they contain protected health information.

2.2 | Setting

The CCHMC neurology clinic has >12 000 annual epilepsy encounters from 6500 unique patients. Of these, there are approximately 100 patients evaluated for resective and nonresective surgical procedures. Surgical referrals for patients with epilepsy are made by neurologists according to International League Against Epilepsy (ILAE) criteria and American Academy of Neurology guidelines.[12,13] Once referred, patients visit an epileptologist, and if indicated, presurgical evaluation proceeds. A surgical review committee, which includes neurologists, epileptologists, neurosurgeons, radiologists, pathologists, and psychologists, reviews the multimodality workup before recommending patients for surgery. This process generally takes 2-6 months.

2.3 | Study population included in the training and test sets

The NLP training set contained positive cases (patients with epilepsy who underwent resective surgical treatment for epilepsy, including lobectomy, corticectomy, and hemispherectomy, corresponding to procedure codes 61510, 61531, 61533-61540, 61542, 61543, 61566, and 61567) and negative controls (nonsurgical patients with epilepsy who were seizure-free for at least 12 months prior to their latest visit). Seizure-free patients were retrospectively identified using structured and unstructured information from the EHR. The training set was updated weekly as patients underwent resective surgery or became seizure-free, resulting in a training cohort that grew

over time. The model was retrained weekly on the most recent training set. Weights in the updated models were not directly impacted by weights in previous models.

All patients with an outpatient neurology clinic visit during the study period (October 31, 2016 to October 30, 2017) were eligible for inclusion. The test set was updated weekly and consisted of all patients with unknown surgical candidacy status. These were patients who (1) had an International Classification of Diseases, Ninth Revision code for epilepsy or convulsions (345.*, 780.3*, and 779.0),[27] (2) had at least one prior office visit in the Division of Neurology, and (3) had a neurology visit scheduled within the next 6 months. Patients with a history of resective epilepsy surgery, a referral for a presurgical evaluation, or documentation of seizure freedom were excluded from the test set. A summary of the inclusion and exclusion criteria is shown in Figure 1.

2.4 | Input to the machine learning model

The input to the NLP application was progress notes that were stored in the EHR. Only notes from visits prior to patients' zero date were used in the training set. The zero date corresponded to either the date of referral for a presurgical evaluation or the last seizure date for surgical and seizure-free patients, respectively. All available neurology progress notes at least 100 words long were used for patients in the test set. No structured data were used. Free text was tokenized into uni-, bi-, and trigrams (words and phrases up to three words long) and converted to lowercase, and all numerals were replaced with the string "_NUM." We mapped any known generic or trade name of antiepileptic drugs (AEDs) to a code representing the medication. For example, lacosamide and Vimpat were both replaced with DRUG_LCM. N-gram frequencies were normalized to Boolean values, under the assumption that the presence or absence of a word or phrase holds significance in a clinical note, and repeated statements in clinical notes do not increase its relevance. This minimized the effect of "copy forwarding." N-grams appearing in only one patient's notes were discarded.

A support vector machine (SVM) classifier was fitted to the training sets using the libsvm implementation in the Python library scikit-learn.[28] SVMs are commonly used in text classification because they perform well with high-dimensional, sparse feature vectors, and are resistant to overfitting. The SVM kernel (linear or Gaussian), C, and gamma parameters were selected via grid search, and the number of n-grams included in the model were selected each week via nested cross-validation. N-gram features were ranked by their discriminative value using a two-sample Kolmogorov-Smirnov test. The number of n-grams included in the model ranged from 2^4 to 2^{15} (16-32 768 ng).

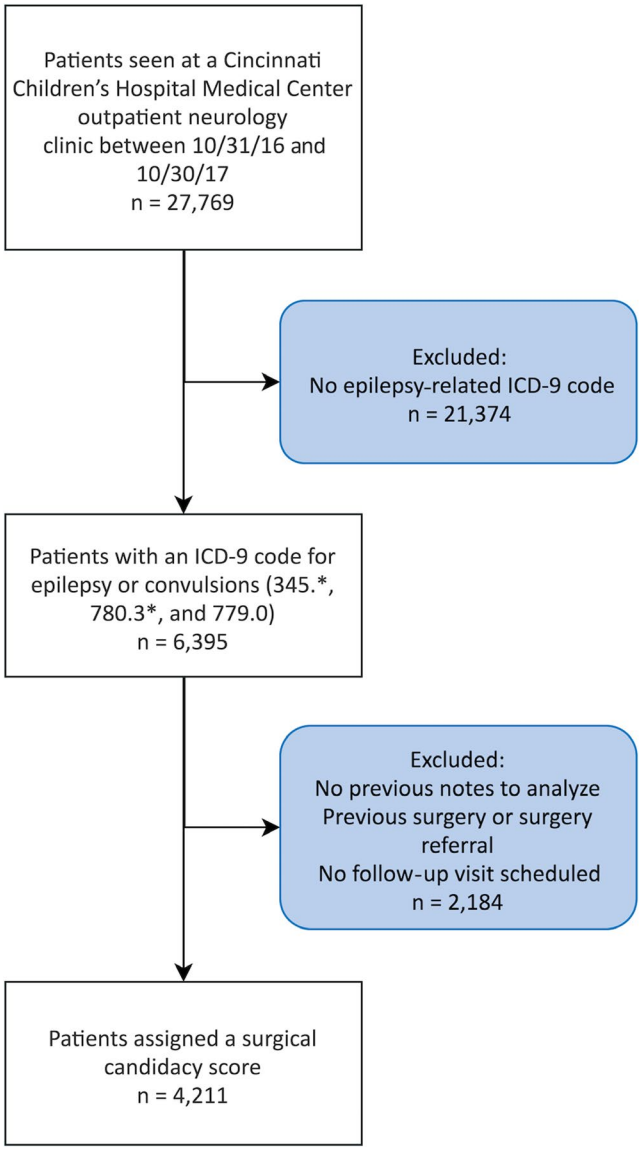

FIGURE 1 Patient inclusion and exclusion from the study. All patients seen between October 31, 2016 and October 30, 2017 at Cincinnati Children's Hospital Medical Center were eligible. ICD-9, International Classification of Diseases, Ninth Revision

2.5 | Patient outcome assessment

Known clinical outcomes (surgery vs seizure freedom without surgery) were used to determine eligibility for inclusion in the training set. For the test set, a random sample of patient charts (some candidates and most noncandidates) were manually reviewed by an epileptologist (H.M.G.). All epilepsy patients, including those who were not evaluated by the application, were eligible to be randomly selected for chart review. This allowed the study team to determine the proportion of potential surgical candidates who were not evaluated by the model. Of the 6395 epilepsy patients who visited a neurology clinic during the study period (Figure 1), 100 (1.6%) were randomly selected for

review. All patients who received a score ≥ 1 the week of their scheduled visit were reviewed by two epileptologists (H.M.G. and K.D.H.-B.). Based on our prior studies, a cutoff score of 1 was considered the threshold for surgical candidacy and was chosen to maximize the positive predictive value (PPV).

For the manual review of patient charts, patients were classified as (1) a potential surgery candidate or (2) not a surgical candidate. Epileptologists defined potential surgical candidates as patients with focal epilepsy who failed at least two adequate AED regimens and were defined as drug-resistant according to ILAE criteria. Patients were not considered candidates if there was clear documentation of seizure freedom >12 months from the last note, or if the seizure type was generalized, undetermined, nonepileptic, or undetermined if epileptic. Patients were not candidates if it was unclear whether there were continued seizures on treatment or if there was a history of noncompliance. All information available in patients' EHR, including notes, medications, procedure reports, and encounters documented before the algorithm scored the patient, was made available to the reviewers. The epileptologists conducted their reviews independently. Discrepancies between surgical candidacy classifications were reconciled by consensus. We only considered patients to be potential surgical candidates if they were classified as candidates by both epileptologists.

2.6 | Output of the model: surgical candidacy scores

Patients evaluated by the model were assigned a surgical candidacy score. Scores roughly followed a standard normal distribution and were centered around zero (range = −5.02 to 3.67). Higher scores indicated greater likelihood of surgery candidacy, and lower scores indicated greater likelihood of seizure freedom.

2.7 | Primary and secondary outcomes

The primary outcome was area under the receiver operating characteristic curve (AUC). Secondary outcomes included the number of surgical candidates that received scores ≥ 1 immediately prior to their visit, interrater agreement of the epileptologists' manual classifications, and other performance metrics.

2.8 | Statistical analysis

Sensitivity, specificity, PPV, negative predictive value (NPV), and number needed to screen were calculated by comparing the NLP-derived surgical candidacy scores to manual classifications from the epileptologists. Number needed to screen was defined as the number of patients with a surgical candidacy score above the cutoff who needed to be reviewed to prevent one delayed or missed referral.[29] The most recent available surgical candidacy scores were used for the randomly selected patients. Ordinary least squares regression was used to test for a trend in the 10-fold cross-validation AUC. Bootstrapping was used to calculate 95% confidence intervals (CIs) for performance measures.[30] Interpretation of Cohen kappa statistic for interrater reliability was 0-0.2 (no agreement), 0.21-0.39 (minimal agreement), 0.4-0.59 (weak agreement), 0.6-0.79 (moderate agreement), 0.8-0.9 (strong agreement), and >0.9 (almost perfect agreement).[31] Analyses were performed using R statistical software version 3.5.1.[32] Two-sided $P < .05$ was considered statistically significant.

3 | RESULTS

3.1 | Clinic population characteristics

There were 27 769 patients who visited an ambulatory neurology clinic over the 1-year study period. Of these, there were 12 019 epilepsy-related visits from 6395 (23%) unique patients. Patient demographics are shown in Table 1. The prevalence of potential candidates for resective surgery in the clinic population, based on chart review of 100 of 6395 (1.6%) randomly selected patients, was 0.07 (95% CI = 0.01-0.12).

3.2 | Application training and performance

The training set was updated weekly and contained a mean of 519 ± 67 patients. The AUC from 10-fold cross-validation averaged 0.90 ± 0.04 (range = 0.83-0.96). The model's cross-validation AUC increased by a mean of 0.002 per week (95% CI = 0.002-0.003, adjusted $R^2 = .81$, $P < .001$) as new patients were added to the training set. The top 25 positively and negatively weighted features derived from the training set are shown in Figure 2. N-grams such as "no seizures," "under excellent control," "spells," and "generalized epilepsy" were among the terms weighted most negatively (ie, poor surgical candidacy). Features that were among the terms weighted most positively (ie, favoring surgical candidacy) included "under suboptimal control," "lack of efficacy," "[number of] seizures," "abnormality," "MRI," and "epilepsy surgery."

Of 6395 epilepsy patients who visited the outpatient clinic during the study period, 4211 (67%) fit inclusion criteria and were evaluated by the model. The mean surgical candidacy

TABLE 1 Characteristics of the patients in the training and test sets

	n (%)				
Variable	**All, n = 6395**	**Scored, n = 4211**	**Reviewed, n = 100**	**Reviewed & scored, n = 58**	**Scored ≥ 1, n = 200**
Age, y					
0-4	1115 (17)	735 (17)	27 (27)	15 (26)	31 (16)
5-9	1491 (23)	1016 (24)	26 (26)	16 (28)	35 (18)
10-14	1525 (24)	1007 (24)	19 (19)	11 (19)	43 (22)
15-17	1001 (16)	650 (15)	8 (8.0)	3 (5.2)	28 (14)
≥18	1263 (20)	803 (19)	20 (20)	13 (22)	63 (32)
Male sex	3481 (54)	2266 (54)	54 (54)	29 (50)	116 (58)
Race					
White	5153 (81)	3384 (80)	84 (84)	52 (90)	169 (85)
Black	712 (11)	484 (11)	8 (8.0)	3 (5.2)	22 (11)
Asian	99 (1.5)	65 (1.5)	0 (0.0)	0 (0)	3 (1.5)
Other	107 (1.7)	65 (1.5)	1 (1.0)	0 (0)	3 (1.5)
Multiracial	230 (3.6)	150 (3.6)	6 (6.0)	2 (3.4)	1 (0.5)
Unknown	94 (1.5)	63 (1.5)	1 (1.0)	1 (1.7)	2 (1.0)
Neurology visits, n					
1	2968 (46)	1371 (33)	52 (52)	20 (34)	29 (15)
2-3	2824 (44)	2404 (57)	38 (38)	31 (53)	136 (68)
4-6	530 (8.3)	406 (9.6)	9 (9.0)	7 (12)	28 (14)
7-10	59 (0.9)	27 (0.6)	1 (1.0)	0 (0)	4 (2.0)
>10	14 (0.2)	3 (0.1)	0 (0.0)	0 (0)	3 (1.5)

Note: Data are shown for all patients with epilepsy who visited the neurology clinic during the study period (All), all patients who satisfied inclusion criteria and were scored by the model (Scored), 100 randomly selected patients from All whose charts were manually reviewed by an epileptologist (Reviewed), patients in both Reviewed and Scored (Reviewed & scored), and patients from "Scored" with a surgical candidacy score ≥ 1 (Scored ≥ 1).

score was −0.28 ± 0.81 (range = −5.02 to 3.67). A representative patient's surgical candidacy scores are provided in Figure 3. At the beginning of the study period, this patient's seizures were well controlled with AEDs. Later, he/she had breakthrough seizures and was referred for a presurgical evaluation. A resectable lesion found previously on magnetic resonance imaging (MRI) was surgically removed shortly after the study period.

3.3 | The application's prospective performance on 100 patients

Figure 4 summarizes the performance of the model on the randomly selected patients. Of the 100 patients, 58 (58%) fit inclusion criteria and were assigned surgical candidacy scores. Five (9%) were potential surgical candidates, and 53 (91%) were noncandidates. There were no differences in age, gender, race, or number of visits between the 58 randomly selected patients and the study population at large ($P > .05$). The model's prospective AUC for these 58 patients was 0.79 (95% CI = 0.62-0.96). The optimal surgical candidacy threshold was a score of 0.22, which yielded a sensitivity of 0.80 (95% CI = 0.28-0.99) and a specificity of 0.77 (95% CI = 0.64-0.88). The probability that a patient receiving a surgical candidacy score ≥ 0.22 was a potential surgical candidate (PPV) was 0.25 (95% CI = 0.07-0.52). The probability that a patient scored <0.22 and was not a candidate (NPV) was 0.98 (95% CI = 0.87-1.00). The number needed to screen was 5.6.

3.4 | Classification of 200 potential surgical candidates

Two-hundred patients received a score ≥ 1. After chart review, 54 (27%) were potential surgical candidates. Of the 54 surgical candidates, 42 (79%) were not referred for a presurgical evaluation during the study period. There was a positive correlation between the number of neurology office visits and the proportion of patients with scores ≥ 1 ($P = .03$). Of the 200 patients, 173 (77%) had between five and 22 neurology visits. The NLP application's classifications were more accurate on patients with more office

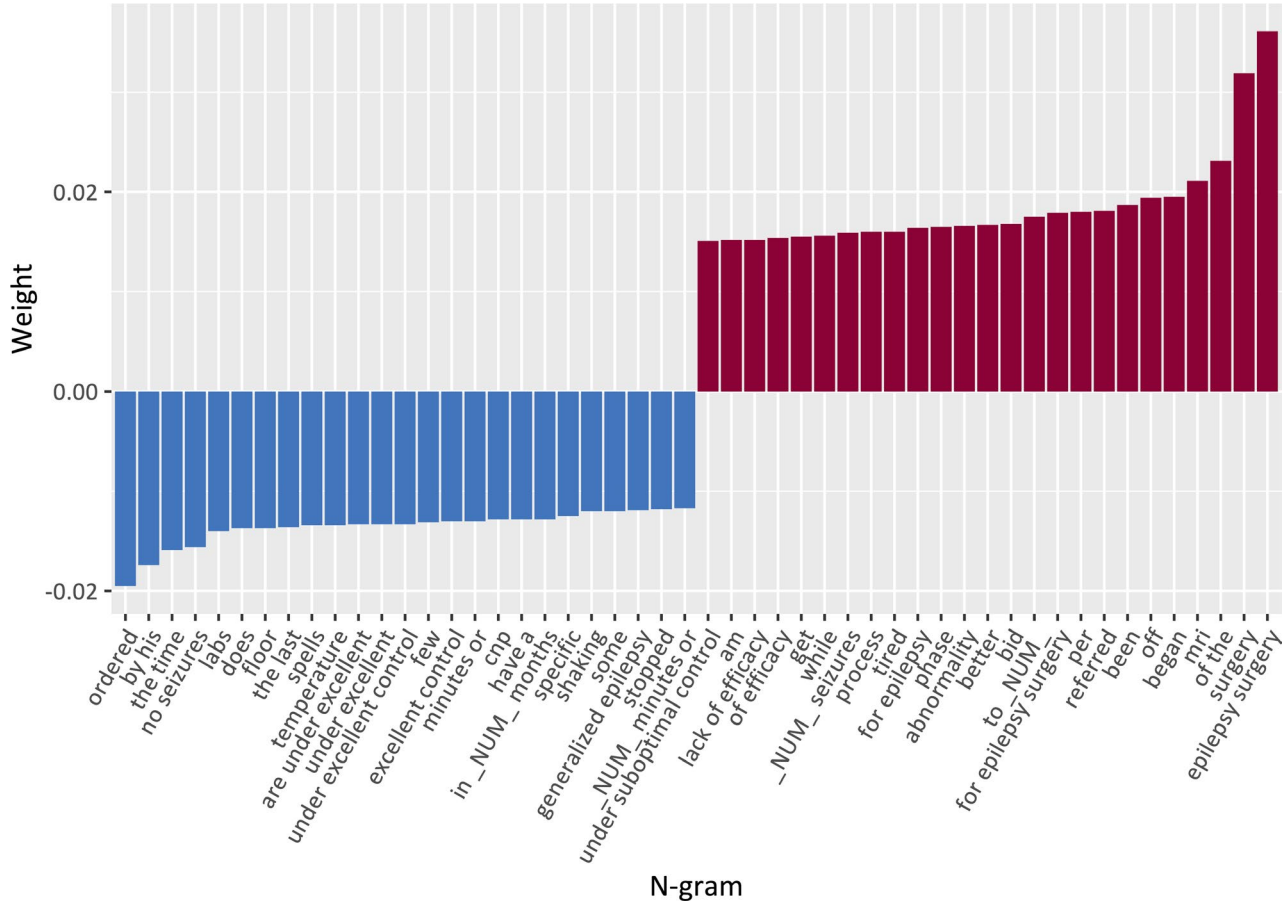

FIGURE 2 Weights of the most heavily weighted features. The feature set was derived from all one-, two-, and three-word phrases contained in a least two different patients' notes. Phrases such as "no seizures" were weighted negatively, whereas phrases such as "under suboptimal control" were weighted positively

visits ($P = .01$). Cohen kappa coefficient for the epileptologists' classifications indicated moderate agreement, at 0.61 (95% CI = 0.49-0.73, $P = .03$).

4 | DISCUSSION

We prospectively evaluated an NLP application that assigned surgical candidacy scores to patients with epilepsy and validated the results through expert chart review. The application evaluated a large number of patients without additional data entry by providers. By updating the model weekly, we are able to present a sustainable approach that incrementally improves accuracy over time.

4.1 | Evaluation of the application

Qualitative inspection of the model's most heavily weighted features revealed that language used by clinicians to indicate uncontrolled epilepsy was generally weighted toward surgical candidacy. Language used to describe controlled epilepsy was weighted toward seizure freedom. Terms to describe seizures that should not be considered for resective surgery, such as "generalized epilepsy" and "spells," the later presumably describing nonepileptic events, were strongly weighted against surgical candidacy. These weights were chosen without incorporating domain knowledge, which allows for continued updating with little additional human effort. This is in contrast to other efforts that hand-engineered a long list of regular expressions to construct epilepsy phenotypes, which would need to be updated over time.[33]

4.2 | Comparison to an existing tool

CASES evaluates the appropriateness of surgical referral using an online calculator. It has high sensitivity (95%),[14] but the majority of patients it identifies (85%) are already being considered for referral by their clinicians.[34] CASES is recommended for patients with focal epilepsy who are >12 years old. In children, it is often not clear whether

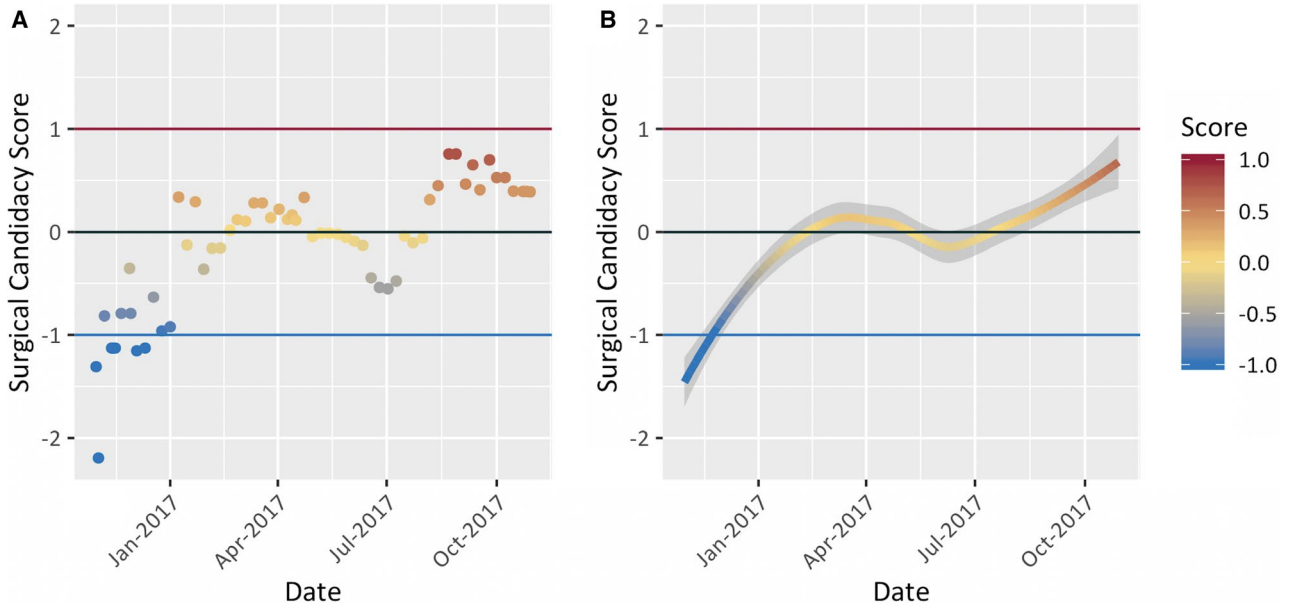

FIGURE 3 Illustration of surgical candidacy scores during the study period (x-axis represents time) for one selected patient. Weekly scores for this patient, who underwent resective epilepsy surgery after the study period, are illustrated in A. Scores varied week to week as the patient accrued notes from more office visits and as the model was continuously updated. The score trend and 95% confidence interval (dark gray) are shown in B. Lower (blue) scores correspond to a higher likelihood of seizure freedom, and higher (red) scores correspond to a higher likelihood that the patient was a candidate for resective epilepsy surgery. This patient was "under good control with monotherapy" early in the study period but had breakthrough seizures throughout the year. The notes indicated that a resectable lesion was discovered on magnetic resonance imaging, and ultimately the patient was referred for a presurgical evaluation and underwent resective surgery shortly after the study period

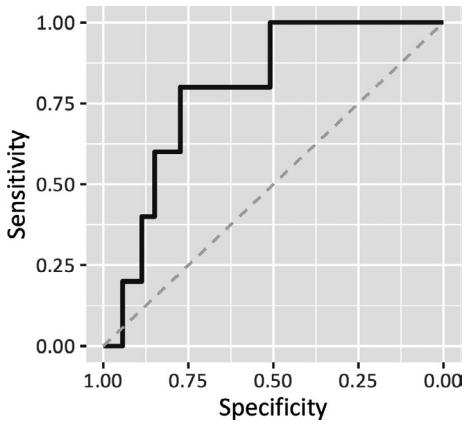

FIGURE 4 Receiver operating characteristic curve for the natural language processing model. A subset of 100 patients with epilepsy who visited a neurology clinic during the study period were randomly selected for chart review. Of these 100 patients, 58 met inclusion criteria and received surgical candidacy scores. These surgical candidacy scores were compared against surgical candidacy classifications from an epileptologist. The area under the curve was 0.79 (95% confidence interval = 0.62-0.96)

epilepsy is focal or multifocal/generalized. This is due to the high proportion of extratemporal epilepsies and lack of detailed patient history (for clues such as aura) in some cases. CASES will require updates and repeat validation if surgical referral guidelines change. In contrast, our model was designed for pediatrics, where multifocal and generalized epilepsies are more common, and its training set is continuously updated. This allows its performance to improve over time and surgical candidacy scores to adapt to evolving clinical practice. Our model can be used without any additional data entry. This facilitates the evaluation of thousands of patients each week.

4.3 | Integrating the application into clinical practice

This application's surgical candidacy scores could be integrated into clinical practice in multiple ways. First, individual patients could be monitored over time to produce a surgical candidacy score trendline, as seen in Figure 3. This could be used to succinctly summarize patients' clinical course, which is known to fluctuate in epilepsy.[35,36] Among patients who undergo resective epilepsy surgery, 26% report a period of at least 1 year of seizure remission prior to surgery.[37] Longitudinal, dynamic surgical candidacy scores would help to visualize current and past surgical candidacy status. Another approach is to alert clinicians before they are scheduled to see a patient with a high probability of surgical candidacy. Alerts have been reported to improve provider performance and patient outcomes when provided under optimal conditions.[20,38] These alerts may enhance provider awareness of surgical candidacy among patients

whom they had previously considered ineligible or for whom had not yet considered recommending surgical treatment.

Because the application produces surgical candidacy scores instead of surgical candidacy classifications, the threshold for a positive screen can be adjusted according to individual clinicians' preference for sensitivity and specificity. The prevalence of potential surgical candidates in this population was 0.07. In this case, high NPV (98%) may be useful for clinicians to rule out whether they should consider referring a patient.

4.4 | Limitations

The results of this study should be viewed within the context of its limitations. First, the interpretability of the model's scores is limited, because the weights of the n-grams are not always intuitive. However, patterns are observable and generally agree with a priori knowledge of criteria for surgical eligibility. The purpose of the tool is not to teach the clinician what features make the patient a good or bad candidate (there are already numerous resources available that thoroughly cover this topic)[39,40]; its purpose is to reliably identify who may be a candidate. Second, patient scores have natural fluctuations over time as the model is continuously updated with new training data. This allows for improved performance over time, but may limit comparing scores from different weeks.

Although NLP may help find patients faster or more comprehensively, its direct effect on surgical outcomes is unknown. Our approach relies on the epilepsy surgery team to determine the potential risk versus benefit of resective surgery for patients identified by the application. Any patient who is recommended by the model must still undergo a full presurgical evaluation, where their potential surgical outcomes would be better defined.

Because the model was trained using patients with known clinical outcomes from one center, the model's classifications could be biased toward the clinical practices at CCHMC. Referral and evaluation of surgical candidacy include subjective criteria that are viewed differently between neurologists, such as the threshold for seizure frequency and severity needed to warrant a surgical recommendation.[17,41] Our chart review was limited by this factor, as evidenced by our moderate interrater reliability, which was comparable to other epilepsy studies.[33] However, our training set is less likely to be influenced by this bias. The model is trained using cases with known clinical outcomes that were dependent upon thorough clinical workups by an interdisciplinary team of epilepsy surgery experts. Incorporating cases from multiple centers would further reduce selection bias. Our previous work suggests that evaluating epilepsy notes across different centers is feasible.[42]

Our model achieved good performance using an SVM classifier here and in earlier studies.[23–25] Recent advances in deep learning[43] and natural language processing,[44] including the use of recurrent neural networks and embedding techniques,[45,46] used here,[47] may represent an avenue of exploration for further improvements in performance.

Given that the model is capable of identifying at least of portion of patients who will undergo future surgery earlier than clinicians, some of the "false positives" will actually become true positives after further follow-up. Thus, we believe the performance we are reporting here may actually be an underestimation of its true performance.

5 | CONCLUSIONS

In conclusion, the NLP model scored patients with epilepsy who could benefit from resective surgery. Its performance was validated prospectively in a clinical setting. Continued use of this model and expansion to additional hospitals and community practices are likely to further improve its generalizability. Future work should include evaluating the effect of alerting physicians of patients' surgical candidacy scores.

ACKNOWLEDGMENTS

We thank the patients who were included in this analysis and the neurologists at CCHMC who have enthusiastically embraced this technology with the hope of improving patient care. This study was funded by a grant from the Agency for Healthcare Research and Quality (AHRQ 1 R21 HS024977-01). Efforts by R.D.S. were supported in part by the National Institutes of Health (K25 HL12595).

CONFLICT OF INTEREST

None of the authors has any conflict of interest to disclose. J.P.P., T.A.G., and H.M.G. report a patent pending on the identification of surgical candidates using natural language processing, licensed to CCHMC. J.P.P. and T.A.G. report a patent pending on processing clinical text with domain-specific spreading activation methods, also licensed to CCHMC. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

AUTHOR CONTRIBUTIONS

B.D.W., H.M.G., J.P.P., and J.W.D. contributed to the conception and design of this study. D.S., J.P.P., and J.W.D. contributed to the acquisition of the data. B.D.W., N.Z., and

R.D.S. performed statistical analyses and interpreted results. B.D.W. and J.W.D. contributed to drafting of the manuscript text. All authors approved the final version of the manuscript.

ORCID

Benjamin D. Wissel https://orcid.org/0000-0002-8967-9296

REFERENCES

1. Kwan P, Brodie MJ. Early identification of refractory epilepsy. N Engl J Med. 2000;342:314–9.
2. Begley CE, Famulari M, Annegers JF, et al. The cost of epilepsy in the United States: an estimate from population-based clinical and survey data. Epilepsia. 2000;41:342–51.
3. Ryvlin P, Rheims S. Epilepsy surgery: eligibility criteria and pre-surgical evaluation. Dialogues Clin Neurosci. 2008;10:91–103.
4. Wiebe S, Blume WT, Girvin JP, Eliasziw M, Effectiveness and Efficiency of Surgery for Temporal Lobe Epilepsy Study Group. A randomized, controlled trial of surgery for temporal-lobe epilepsy. N Engl J Med. 2001;345:311–8.
5. Wyllie E, Comair YG, Kotagal P, Bulacio J, Bingaman W, Ruggieri P. Seizure outcome after epilepsy surgery in children and adolescents. Ann Neurol. 1998;44:740–8.
6. Dwivedi R, Ramanujam B, Chandra PS, et al. Surgery for drug-resistant epilepsy in children. N Engl J Med. 2017;377:1639–47.
7. Moseley BD, Nickels K, Wirrell EC. Surgical outcomes for intractable epilepsy in children with epileptic spasms. J Child Neurol. 2012;27:713–20.
8. Paolicchi JM, Jayakar P, Dean P, et al. Predictors of outcome in pediatric epilepsy surgery. Neurology. 2000;54:642–7.
9. Engel J Jr, McDermott MP, Wiebe S, et al. Early surgical therapy for drug-resistant temporal lobe epilepsy: a randomized trial. JAMA. 2012;307:922–30.
10. Skirrow C, Cross JH, Cormack F, Harkness W, Vargha-Khadem F, Baldeweg T. Long-term intellectual outcome after temporal lobe surgery in childhood. Neurology. 2011;76:1330–7.
11. Jobst BC, Cascino GD. Resective epilepsy surgery for drug-resistant focal epilepsy: a review. JAMA. 2015;313:285–93.
12. Engel J Jr, Wiebe S, French J, et al. Practice parameter: temporal lobe and localized neocortical resections for epilepsy: report of the Quality Standards Subcommittee of the American Academy of Neurology, in association with the American Epilepsy Society and the American Association of Neurological Surgeons. Epilepsia. 2003;44:741–51.
13. Kwan P, Arzimanoglou A, Berg AT, et al. Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia. 2010;51:1069–77.
14. Lukmanji S, Altura KC, Rydenhag B, Malmgren K, Wiebe S, Jetté N. Accuracy of an online tool to assess appropriateness for an epilepsy surgery evaluation—a population-based Swedish study. Epilepsy Res. 2018;145:140–4.
15. Haneef Z, Stern J, Dewar S, Engel J Jr. Referral pattern for epilepsy surgery after evidence-based recommendations: a retrospective study. Neurology. 2010;75:699–704.
16. Roberts JI, Hrazdil C, Wiebe S, et al. Neurologists' knowledge of and attitudes toward epilepsy surgery: a national survey. Neurology. 2015;84:159–66.
17. Steinbrenner M, Kowski AB, Holtkamp M. Referral to evaluation for epilepsy surgery: reluctance by epileptologists and patients. Epilepsia. 2019;60:211–9.
18. Roberts JI, Hrazdil C, Wiebe S, et al. Feasibility of using an online tool to assess appropriateness for an epilepsy surgery evaluation. Neurology. 2014;83:913–9.
19. Jette N, Quan H, Tellez-Zenteno JF, et al. Development of an online tool to determine appropriateness for an epilepsy surgery evaluation. Neurology. 2012;79:1084–93.
20. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ. 2005;330:765.
21. Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293:1223–38.
22. Pestian J, Deleger L, Savova G, Dexheimer JW, Solti I. Natural language processing—the basics. In: Hutton JJ, editor. Pediatric biomedical informatics: computer applications in pediatric research. New York: Springer, 2012; p. 149–72.
23. Cohen KB, Glass B, Greiner HM, et al. Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning. Biomed Inform Insights. 2016;8:11–8.
24. Matykiewicz P, Cohen K, Holland KD, et al. Earlier identification of epilepsy surgery candidates using natural language processing. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. Sofia: Association for Computational Linguistics, 2013; p. 1–9.
25. Wissel BD, Greiner HM, Glauser TA, et al. Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. Epilepsia. 2019;60:e93–8.
26. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1–73.
27. Jetté N, Reid AY, Quan H, Hill MD, Wiebe S. How accurate is ICD coding for epilepsy? Epilepsia. 2010;51:62–9.
28. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
29. Rembold CM. Number needed to screen: development of a statistic for disease screening. BMJ. 1998;317:307–12.
30. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Stat Sci. 1996;11:189–228.
31. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22:276–82.
32. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
33. Barbour K, Hesdorffer DC, Tian N, et al. Automated detection of sudden unexpected death in epilepsy risk factors in electronic medical records using natural language processing. Epilepsia. 2019;60:1209–20.
34. Roberts JI, Hrazdil C, Wiebe S, et al. Neurologists' knowledge of and attitudes toward epilepsy surgery: a national survey. Neurology. 2015;84:159–66.
35. Berg AT, Levy SR, Testa FM, D'Souza R. Remission of epilepsy after two drug failures in children: a prospective study. Ann Neurol. 2009;65:510–9.

36. Callaghan B, Schlesinger M, Rodemer W, et al. Remission and relapse in a drug-resistant epilepsy population followed prospectively. Epilepsia. 2011;52:619–26.
37. Berg AT, Langfitt J, Shinnar S, et al. How long does it take for partial epilepsy to become intractable? Neurology. 2003;60:186–90.
38. Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293:1223–38.
39. Cross JH, Jayakar P, Nordli D, et al. Proposed criteria for referral and evaluation of children for epilepsy surgery: recommendations of the Subcommission for Pediatric Epilepsy Surgery. Epilepsia. 2006;47:952–9.
40. Ryvlin P, Cross JH, Rheims S. Epilepsy surgery in children and adults. Lancet Neurol. 2014;13:1114–26.
41. Fois C, Kovac S, Khalil A, et al. Predictors for being offered epilepsy surgery: 5-year experience of a tertiary referral centre. J Neurol Neurosurg Psychiatry. 2016;87:209–11.
42. Connolly B, Matykiewicz P, Bretonnel Cohen K, et al. Assessing the similarity of surface linguistic features related to epilepsy across pediatric hospitals. J Am Med Inform Assoc. 2014;21:866–70.
43. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
44. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015;349:261–6.
45. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014; p. 1532–43.
46. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in neural information processing systems 26. Red Hook, NY: Curran Associates, Inc., 2013; p. 3111–9.
47. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572:116–9.

How to cite this article: Wissel BD, Greiner HM, Glauser TA, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. *Epilepsia*. 2019;00:1–10. https://doi.org/10.1111/epi.16398