Received: 16 December 2022 | Revised: 26 April 2023 | Accepted: 26 April 2023

DOI: 10.1111/epi.17633

Epilepsia™

RESEARCH ARTICLE

Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes

Kevin Xie[1,2] | **Ryan S. Gallagher**[2,3] | **Russell T. Shinohara**[4,5] | **Sharon X. Xie**[6] | **Chloe E. Hill**[7] | **Erin C. Conrad**[2,3] | **Kathryn A. Davis**[2,3] | **Dan Roth**[8] | **Brian Litt**[1,2,3] | **Colin A. Ellis**[2,3]

[1]Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Center for Neuroengineering and Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[4]Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[5]Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[6]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[7]Department of Neurology, University of Michigan, Ann Arbor, Michigan, USA

[8]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence
Colin A. Ellis, Hospital of the University of Pennsylvania, 3400 Spruce St., 3 West Gates Building, Philadelphia, PA 19104, USA.
Email: colin.ellis@pennmedicine.upenn.edu

Funding information
American Academy of Neurology; Jonathan and Bonnie Rothberg; Mirowski Family Foundation; National Institute of Neurological Disorders and Stroke, Grant/Award Number: 1DP1 OD029758 and K23NS121520; Office of Naval Research, Grant/Award Number: N00014-19-1-2620

Abstract

Objective: Electronic medical records allow for retrospective clinical research with large patient cohorts. However, epilepsy outcomes are often contained in free text notes that are difficult to mine. We recently developed and validated novel natural language processing (NLP) algorithms to automatically extract key epilepsy outcome measures from clinic notes. In this study, we assessed the feasibility of extracting these measures to study the natural history of epilepsy at our center.

Methods: We applied our previously validated NLP algorithms to extract seizure freedom, seizure frequency, and date of most recent seizure from outpatient visits at our epilepsy center from 2010 to 2022. We examined the dynamics of seizure outcomes over time using Markov model-based probability and Kaplan–Meier analyses.

Results: Performance of our algorithms on classifying seizure freedom was comparable to that of human reviewers (algorithm F_1 = .88 vs. human annotator κ = .86). We extracted seizure outcome data from 55 630 clinic notes from 9510 unique patients written by 53 unique authors. Of these, 30% were classified as seizure-free since the last visit, 48% of non-seizure-free visits contained a quantifiable seizure frequency, and 47% of all visits contained the date of most recent seizure occurrence. Among patients with at least five visits, the probabilities of seizure freedom at the next visit ranged from 12% to 80% in patients having

seizures or seizure-free at the prior three visits, respectively. Only 25% of patients who were seizure-free for 6 months remained seizure-free after 10 years.

Significance: Our findings demonstrate that epilepsy outcome measures can be extracted accurately from unstructured clinical note text using NLP. At our tertiary center, the disease course often followed a remitting and relapsing pattern. This method represents a powerful new tool for clinical research with many potential uses and extensions to other clinical questions.

KEYWORDS

clinical informatics, electronic health record, seizure freedom, seizure frequency

1 | INTRODUCTION

The electronic health record (EHR) contains large amounts of free text data, including clinically meaningful outcome measures, across all areas of medicine. At epilepsy centers, these measures include seizure frequency, seizure freedom, and the date of most recent seizure, often present in the text of progress notes. For epilepsy, seizure freedom is a critical outcome measure, and documentation of seizure frequency and time since most recent seizure recorded are among the American Academy of Neurology's epilepsy quality measures.[1,2] Presently, these specific outcome measures are represented in unstructured text data in a myriad of formats, precluding traditional text mining approaches.[2,3] Automated processes to accurately extract seizure frequency and freedom from EHRs would permit important research, such as comparisons of effectiveness of treatment interventions, retrospective clinical trials, and natural history studies, all with the potential to improve the delivery of care for patients with epilepsy. These same principles apply in all medical specialties with their own disease-specific outcome measures.

We recently developed and validated a natural language processing (NLP) algorithm to extract seizure freedom, seizure frequency, and date of most recent seizure from the text of outpatient progress notes for patients with epilepsy.[4,5] Using annotated clinical notes, we fine-tuned and applied state-of-the-art transformer language models to rapidly read and comprehend clinical note text. The algorithm achieved near-human performance at classifying patients as seizure-free at each clinic visit (median accuracy = 84%) and human performance at determining seizure frequency (accuracy = 88%, F_1 score = 85%) and the date of most recent seizure (accuracy = 86%, F_1 score = 83%).[4,5]

In this study, we sought to (1) determine the feasibility of extracting these epilepsy outcomes measures from an EHR at large scale and (2) use this approach to characterize epilepsy outcomes over time at our academic epilepsy center. We applied our method to 55 630 outpatient progress notes from the epilepsy center at our institution over a 12-year period. We examined patterns of seizure outcomes over time, and estimated probabilities of future seizure freedom based on past seizure freedom. We also analyzed patterns of model errors with an eye toward future improvements.

Key Points

- Epilepsy outcome measures can be extracted accurately and at scale from unstructured clinical note text using NLP
- The disease course at our single academic epilepsy center often followed a remitting and relapsing pattern rather than a simple dichotomy of drug-responsive versus drug-resistant epilepsy
- This method represents a powerful new tool for clinical research with many potential uses and extensions to other clinical questions

2 | MATERIALS AND METHODS

2.1 | Data collection

This research was approved by the institutional review board of the University of Pennsylvania with a waiver of informed consent.

We identified all outpatient visits from years 2010 through 2022 for patients with epilepsy-related International Classification of Diseases (ICD) codes 345 and 780.3 (ICD-9), and G40 and R56 (ICD-10) who were seen at our comprehensive epilepsy center. From the EHRs, we extracted the full progress note text, author, and visit date. We filtered for notes written by epileptologists

and epilepsy nurse practitioners, and excluded attending attestations, addendums, and notes fewer than five lines in length.

2.2 | NLP model development and implementation

Our methods for implementing our NLP algorithm have been described elsewhere in detail[4,5] and are summarized in Figure 1. Briefly, we fine-tuned five seeds of pretrained Bio_ClinicalBERT and RoBERTa transformer language models on human-annotated clinical notes.[4,6,7] Transformer models are deep-learning-based neural networks trained to understand the meanings of words and the relationships between them.[8,9] These models classified patients as seizure-free, and extracted the span of text that contained their seizure frequency and/or date of last seizure.[4] We defined each visit as "seizure-free" if the patient had not had seizures since their last visit or within the past 1 year, whichever was more recent. These classifications are not static classifications of the patient, but rather are classifications of the patient at each of their visits for the defined period between visits. Our current algorithms did not distinguish between different types of seizures. We then used a combination of neural summarization with the T5 language model[10] and custom rules-based quantification to convert the extracted text spans into quantitative frequencies and date–time objects, respectively.[5]

In this study, we applied this pipeline to our dataset of outpatient office notes. For each note, we classified patients as seizure-free or having recent seizures, and we quantified the seizure frequency and date of most recent seizure. To improve performance, we repeated this process five times, once for each seed of the pipeline, creating five independent sets of predictions. We merged these predictions through plurality voting to generate a final set of predictions, a method that has been shown in other contexts to improve the performance of NLP models.[11,12]

2.3 | Statistical analysis

Seizure outcomes were described with descriptive statistics. Because the primary unit of our analysis was office visits, and because the pattern of office visits may be affected by seizure outcome status (e.g., seizure-free patients may have less frequent and fewer total visits than patients with ongoing seizures), we calculated the median times to next visit following a seizure-free or not-seizure-free visit across patients with at least five visits. Furthermore, we compared the frequency of visits, total number of visits, and total time at our institution between those patients who were nearly always seizure-free (at least 80% of visits seizure-free) and those patients who were rarely or never seizure-free (at least 80% of visits not seizure-free) using two-sided Mann–Whitney *U*-tests.

To estimate the probability of seizure freedom at a given visit based on prior visits, we used Markov model-based chains, where the nodes of the chain represent the patients' state (having recent seizures or seizure-free) at each visit, and the probability of seizure freedom is dependent on the previous nodes. We included all patients with at least five visits, including at least four consecutive visits without missing classifications. We calculated transition probabilities by counting the number of occurrences of

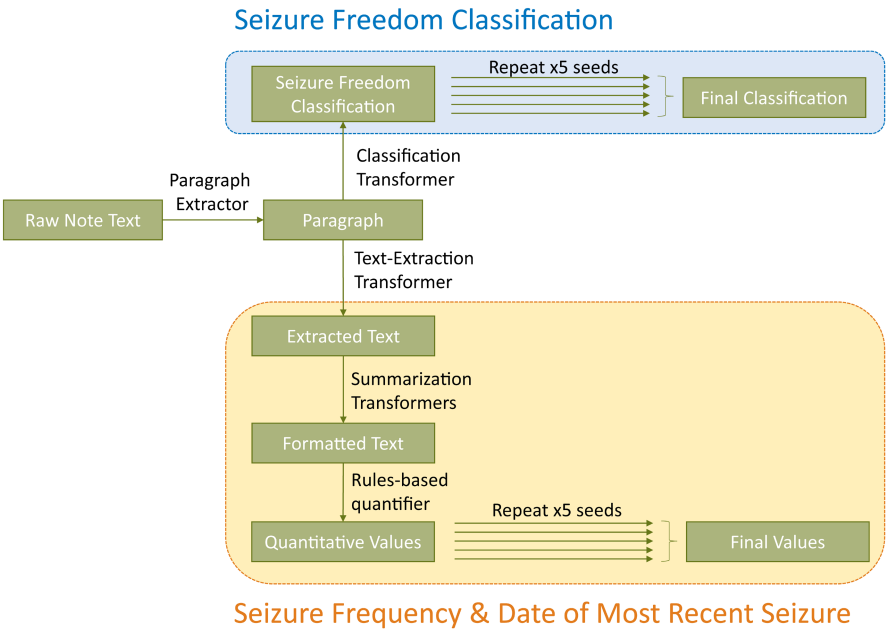

FIGURE 1 Schematic methods of natural language processing pipeline. Seizure freedom was approached as a classification task. Seizure frequency and date of most recent seizure were text extraction tasks, followed by additional summarization and quantification steps to determine final frequency and date values. The algorithm was repeated with five different seeds and used plurality voting to arrive at the final output. Transformer models are a class of deep-learning-based neural networks that are trained on large amounts of data to understand and reproduce human language.

each possible sequence of states across the cohort. We calculated 95% confidence intervals (CIs) for these probabilities using nonparametric bootstrapping by subject with 10 000 iterations.

To further analyze how seizure outcome changed over time, we performed a Kaplan–Meier time-to-event analysis. We identified patients with at least three visits and found their first interval that spanned at least 6 months of visits ("baseline interval"), excluding visits with missing classifications. Patients then entered the time-to-event analysis starting from the end of this baseline interval (time 0). At the time of entry to the time-to-event analysis, we classified patients as "seizure-free" if all visits during the baseline interval were classified as seizure-free. We followed each patient across all remaining visits following entry into the time-to-event analysis. Patients with <1 year of data after the baseline interval were excluded. We monitored over time for the first visit with breakthrough seizures. Patients were removed from the analysis (censored) after their last available visit. To identify potential bias in our time-to-event analysis, we checked whether there was a significant correlation between length of follow-up and proportion of visits with recent seizures for these patients. We further compared the age, gender, and race of censored and noncensored patients to determine whether there were differences between these patients using two-sided Mann–Whitney *U*-tests and chi-squared tests.

We validated our plurality voting method on classifying seizure freedom by comparing its classification performance (F_1) against the agreement (Cohen κ) of our annotators on the same task, and on extracting seizure frequencies and dates of last seizure by comparing its "agreement" with the gold standard values to our human annotators' "agreement" on the same task. Our human annotations and annotators were those from our previous validation study.[4] Gold standard seizure frequencies and dates of last seizures were generated by merging and adjudicating human annotations (see our previous study for details[4]); as such, human agreement with the gold standard is inflated. In our previous study, human annotations (and the adjudicated gold standard annotations) were text strings;[4] here, we converted these human-annotated text strings to quantitative frequency and date–time values using the same summarization and quantification steps described above.

Specifically for seizure frequency and date of last seizure, we defined "agreement" as the number of examples where the algorithm or human annotators matched the gold standard divided by the total number of examples. Matches were defined by both (1) correctly identifying when a seizure frequency or date of last seizure was present within the medical note and (2) correctly returning the seizure frequency or date of last seizure value when it did exist. We used this custom agreement metric over established metrics because the extraction of seizure frequency and date of last seizure must be divided into two simultaneous processes. The first categorically determines whether either measure exists within the note (akin to a classification task), whereas the second finds the correct numerical value when it does exist on a continuous domain (a quantification task). No single metric can appropriately handle these two processes at once. For example, the F_1 metric can measure categorical performance on the classification process, but cannot handle continuous performance (e.g., is a prediction of "12 sz/mo" vs. a correct answer of "3 sz/mo" a false positive or negative?). Similarly, metrics like intraclass correlation or root mean square error measure continuous performance on the quantification process, but fail on categorical performance (e.g., what is the error between "no answer" and "3 sz/mo"?). In contrast, our definition of "agreement" concisely captures both processes simultaneously.

For rigor, we include positive predictive value and sensitivity on classifying seizure freedom in Table S2A. For extracting seizure freedom and date of last seizure, we also include the F_1 score and Fisher's one-way random intraclass correlation (ICC) for the classification and quantification processes, respectively, in Table S2B.[13] We found the one-way random ICC sufficient in comparison to other ICCs that account for potential bias according to the selection criteria outlined in Liljequist et al.[14]

All analyses were performed using Python software including packages numpy, pandas, pingouin, lifelines, and scipy. We include links to our NLP models and code in our Data Availability Statement.

3 | RESULTS

3.1 | Cohort

We analyzed 55 630 notes from 9510 patients. Notes were written by 53 unique authors. Notes included 5725 (10%) new patient visits and 49 905 (90%) follow-up visits. Of the 9510 unique patients, 7036 (74%) were seen more than once for 53 156 total visits, and 3682 patients (39%) were seen at least five times for 43 999 total visits. Demographic data were available for 8741 patients (92% of entire cohort; Table S1).

3.2 | Seizure outcomes

Outcome classifications for the 55 630 visits are shown in Table 1. The minority of all visits were classified as seizure-free (16 688 visits, 30%).

Among the 9510 unique patients, 718 (8%) were seizure-free at every visit, 3572 (38%) were having recent seizures at every visit, 4956 (52%) had some visits of each type, and 264 patients (3%) had only unclassified visits. Among the 3682 patients with at least five visits (Figure 2), 110 (3%) were seizure-free at every visit, 391 (11%) were having seizures at every visit, and 3084 (84%) had some visits of both types.

Across all visits, the median time to next visit was 4.1 months (interquartile range [IQR] = 2.1–6.8 months). After a seizure-free visit, the median time to next visit was 6.0 months (IQR = 3.5–11.3 months); after a visit with recent seizures, the median time to next visit was 3.5 months (IQR = 1.8–6.1 months). There were 564 patients who were classified as seizure-free at >80% of their visits and 1021 patients who were classified as not seizure-free at >80% of their visits. The more seizure-free group had less frequent visits (median = 1.6 vs. 2.9 visits/year, $p < .001$) and fewer total visits (median = 8 vs. 10 visits, $p < .001$), and spent more total time at our institution (median time between first and last visit = 69 months vs. 43 months, $p < .001$).

There were 2122 patients with new patient visits classified as having recent seizures and at least three more follow-up visits. Of these patients, 710 (33%) achieved at least 6 months of seizure freedom at the time of last follow-up.

Seizure frequency was extracted from 16 404 visits (48% of visits classified as having recent seizures), and the date of most recent seizure was extracted from 26 098 visits (47% of all notes). These values are similar to our prior manual annotation of a subset of notes,[4] and therefore reflect the expected rate at which these outcome measures are stated in our office notes, rather than failures of the algorithms. When a seizure frequency was detected, the median seizure frequency was three seizures per month. When a date of most recent seizure was detected, these dates ranged from 0 days before the visit to 60 years before the visit (median = 4.5 months before the visit).

TABLE 1 Outcome classifications.

Outcome	*n* (%) of visits
Entire cohort	55 630 (100%)
Seizure freedom classification	
Seizure-free	16 688 (30%)
Having recent seizures	34 452 (62%)
Unclassified	4490 (8%)
Seizure frequency extracted	16 404 (48%[a])
Date of most recent seizure extracted	26 098 (47%)

[a]Seizure frequency expressed as percentage of visits classified as having seizures.

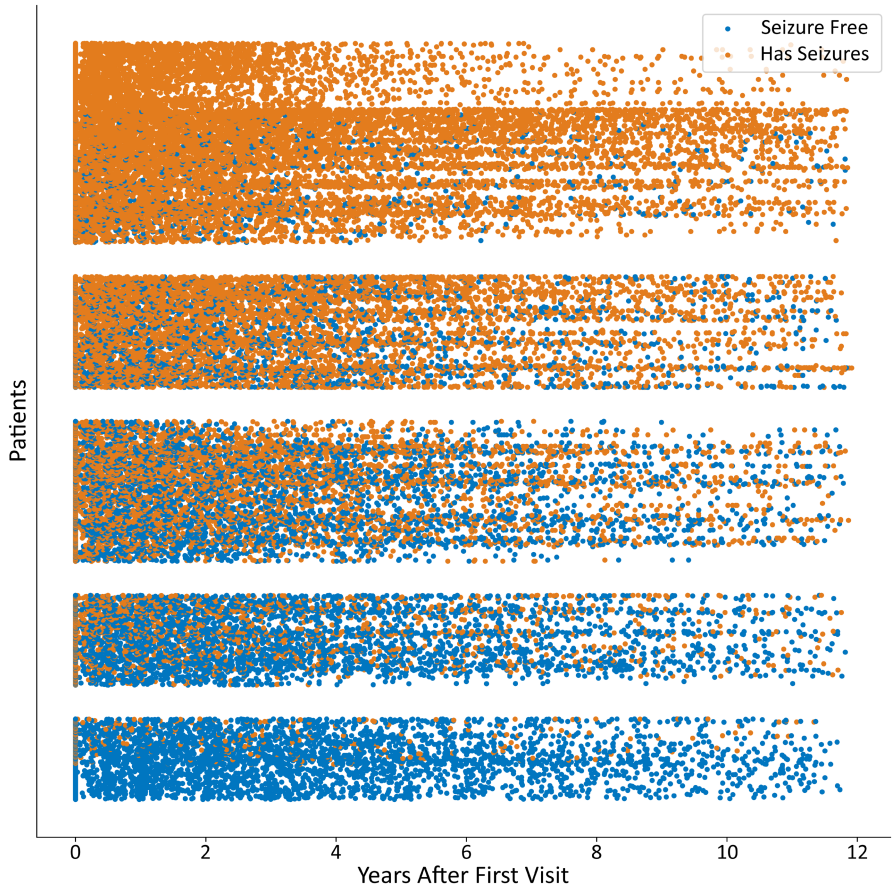

FIGURE 2 Seizure freedom classifications over time. A visual summary represents our large patient cohort and the relapsing–remitting nature of epilepsy. Shown are 43 999 clinic visits (dots) from 3682 patients (rows) seen at least five times at our institution from 2010 to 2022. Each visit was classified by the natural language processing algorithm as seizure-free or having recent seizures. Visits with unknown classification are not shown. Patients were sorted based on fraction of visits that were seizure-free, stratified into quintiles of proportion of visits classified as seizure-free (0–20% of visits, 21–40%, etc.), with white space between each quintile.

3.3 | Estimating probabilities of seizure freedom

We used a third-order Markov model-based approach to estimate the probability of seizure freedom at a given visit based on the outcome classification of the previous three visits (Figure 3). This analysis included 21 638 visits from 3308 patients. Patients who were seizure-free at all three previous visits had a seizure freedom probability of 80% (95% CI = 79%–82%) at the next visit. At the other extreme, patients who were having seizures at all three previous visits had a seizure-freedom probability of 12% (95% CI = 12%–13%) at the next visit. Among patients having seizures at only one of the previous three visits, the probability of seizure freedom at the next visit ranged from 46% to 67%, and varied in a stepwise fashion depending on whether the preceding visit with seizures was the first, second, or third most recent visit.

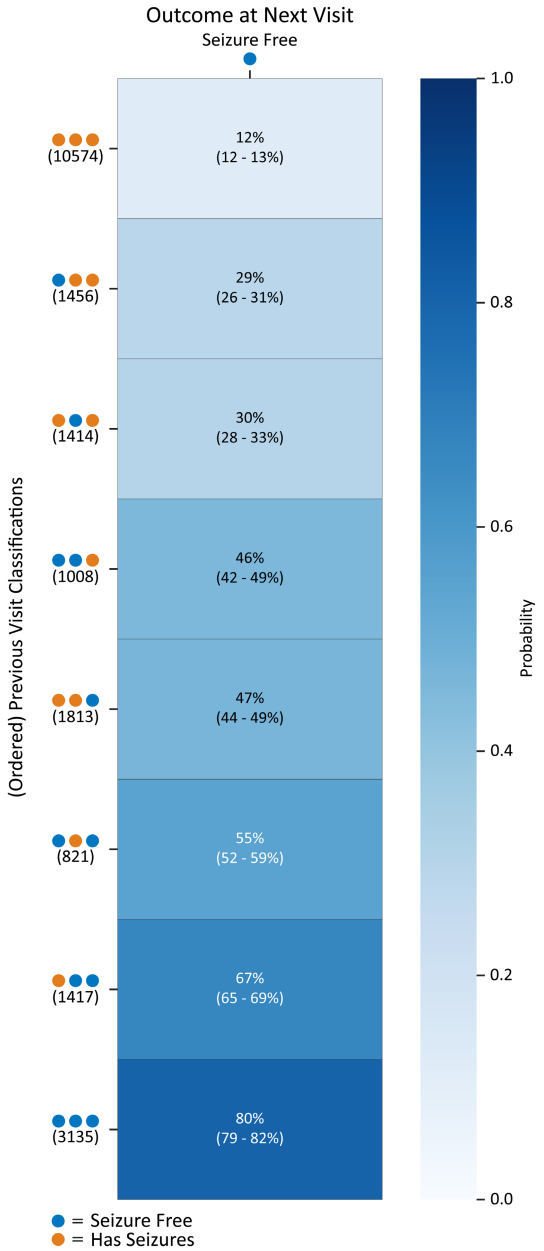

FIGURE 3 Probability of seizure freedom based on the preceding three visits. Probabilities were calculated using a third-order Markov-like model with 95% confidence intervals. Y-axis markers denote the order of classifications in the three previous visits, with sample sizes in parentheses. Darker colors indicate higher probability.

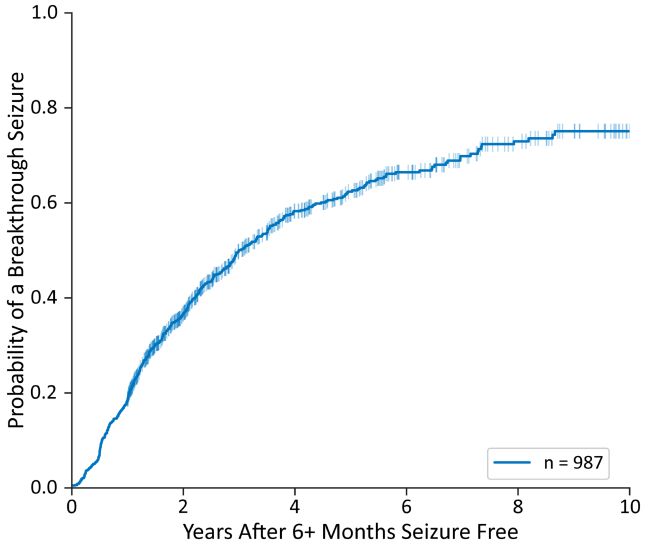

FIGURE 4 Kaplan–Meier time-to-event analysis. Patients who were seizure-free during the 6-month baseline interval were monitored from the end of the baseline interval (time 0) for breakthrough seizures. Patients with <1 year of follow-up from the end of the baseline interval were excluded. Censoring is indicated with a vertical tick.

For our Kaplan–Meier time-to-event analysis, there were 987 patients who were seizure-free during a 6-month baseline interval and had at least 1 year of follow-up. In this cohort, 50% of patients had a breakthrough seizure within 3 years, and 75% had a breakthrough seizure by 10 years (Figure 4). There was no correlation between proportion of visits with seizures and duration of follow-up (Spearman correlation = .013, $p = .69$). Additionally, we found no significant differences in gender, age, or race between the censored and uncensored patients (all $p > .05$).

3.4 | Accuracy and quality control of NLP models

The accuracy of our current NLP methods compared to human reviewers has been previously described.[4] In this study, we observed that plurality voting across our models increased model performance (Figure 5). For classification of seizure freedom, our current model had .88 F_1; for comparison, the average interrater agreement between

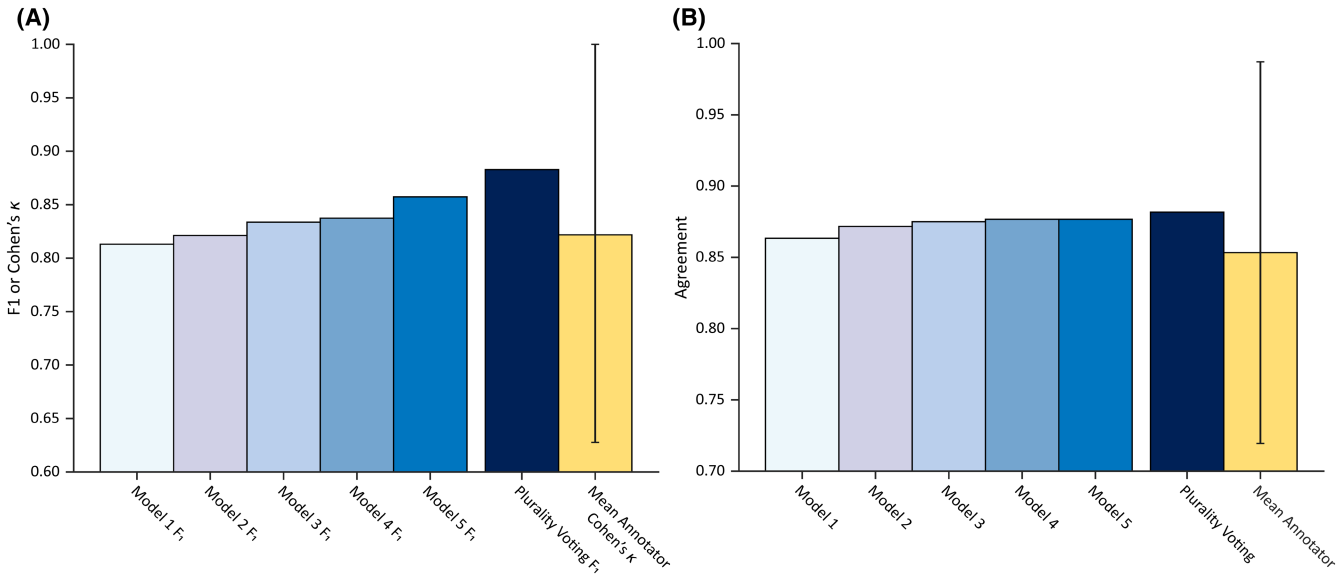

FIGURE 5 Accuracy of the natural language processing models. (A) Classification of visits as seizure-free or having recent seizures. Models 1–5 and final result of plurality voting were measured as F_1 compared to the ground truth annotations. Human performance was measured as Cohen's κ of 15 human reviewers (mean ± 95% confidence interval). (B) Quantifying seizure frequency and date of most recent seizure (combined for this analysis). Agreement of both model and human were measured in comparison to ground truth quantitative values (frequency value or date).

human reviewers was .82 Cohen κ (95% CI = .63–1.00). For seizure frequency and date of last seizure (combined for this analysis), our current model had 88% agreement with the gold standard, compared to 85% (95% CI = 72%–99%) for human reviewers. Additional metrics of model performance are in Tables S2A,B.

Next, we identified extreme values of seizure frequency and date of most recent seizure to determine whether these represented algorithm errors. The 10 highest seizure frequency values ranged from 1521 to 21 900 seizures per month. We manually verified that seven of 10 of these extreme values were correct interpretations of the original note text (Table S3). For date of most recent seizure, a total of 55 notes (.1%) returned dates after the visit date, that is, in the future. We manually reviewed five of these and found that four of five were correct interpretations of the original text, that is, the note erroneously documented a future date of the patient's seizure (Table S4). We also manually reviewed the five notes with the earliest dates of most recent seizure (years ranging 1960–1972) and found that all five were correct in older patients who had been seizure-free for decades (Table S5).

Finally, we identified "contradictory" model outputs: visits classified as seizure-free and with a nonzero seizure frequency (826 visits, 1.5%), visits classified as seizure-free and with a date of most recent seizure after the last visit or within the past 1 year (642 visits, 1.2%), or both (90 visits, 0.2%). We manually reviewed 200 of these contradictory notes, randomly selected. The most common cause of contradiction (59/200, 30%) was multiple seizure types with conflicting information. For example, a note might say, "Since last visit, no convulsions. Isolated auras occur twice per month." The NLP model might classify this note as seizure-free, based on the phrase "no convulsions," while also reporting a seizure frequency of "twice per month." Other causes of contradictory model outputs were failure to distinguish current from outdated information, often due to copy-forwarding or summarizing of old information in note text; extracting the frequency of nonseizure events, such as headaches; and contradictions within the original note (Table S6).

4 | DISCUSSION

In this study, we used a novel NLP algorithm to extract seizure freedom, seizure frequency, and date of most recent seizure from 55 630 free text clinic notes from 9510 unique patients with epilepsy, written by 53 unique authors at a single academic medical center. The algorithm performed these tasks with accuracy comparable to human reviewers. The majority of visits were not seizure-free (62%); the majority of patients (84% of those with five or more visits) had a mix of visits with and without recent seizures. Seizure freedom probability at next visit could be estimated from the preceding three visits. Of patients seizure-free for at least 6 months, 50% had relapsed by 3 years and 75% had relapsed by 10 years.

A tool that accurately extracts clinically meaningful outcome measures from note text has many potential

research applications. Extracting epilepsy outcomes from EHRs has been a major challenge requiring manual review by human readers. Prior efforts to extract epilepsy-related information using NLP have been limited to traditional machine learning and basic rules-based approaches that work only in limited contexts[3,15]; furthermore, few have extracted critical outcome measures like seizure frequency, freedom, and time since last seizure.[3,16] For example, Fonferko-Shadrach and colleagues[17] developed an algorithm that used a combination of rules-based and statistical techniques to extract a number of epilepsy-related variables including seizure frequency, with an F_1 score of .66. Decker and colleagues[18] used a rules-based algorithm to extract seizure frequencies from note text, with an F_1 score of .82 on an internal test set and .40 on an external test set. None of these methods extracted seizure freedom as a distinct outcome measure. Our models extracted seizure freedom, seizure frequency, and date of most recent seizure with accuracies comparable to human readers. In the current study, we demonstrated incremental improvements in the model compared to our prior reports,[4,5] and we analyzed the model's errors with an eye toward continued improvements. Additionally, because our methods are based on Google AI's transformer models, they will be easily adaptable to a wide range of research questions, in epilepsy or other disorders.

Studying outcome measures over time is important for understanding the natural history and prognosis of epilepsy. The landmark study of Kwan and Brodie[19] found that two thirds of patients were seizure-free, defined as no seizures for ≥1 year at time of last follow-up. Other population-based epidemiological studies have found that, if followed long enough, the majority of persons with epilepsy achieved terminal remission, defined as no seizures for ≥5 years at last follow-up.[20–23] However, charting patients' outcomes over time reveals a more complex and dynamic course than these simple binaries suggest. Epidemiological studies have found that fewer than one quarter of patients have early and sustained remission after epilepsy diagnosis.[22,24] For many patients, a remitting–relapsing course is common, with periods of 1 or more years of seizure freedom, interrupted by breakthrough seizures, sometimes repeating this pattern multiple times.[25–30] The risk of relapse decreases with longer duration of remission, but even in patients with >10 years of remission, a substantial proportion will relapse.[28] The converse may also be true. One study of drug-resistant epilepsy, defined as seizures at least monthly despite two or more antiseizure medications, found that one third of those patients not undergoing epilepsy surgery achieved at least 1 year of seizure freedom at some point during 7 years of follow-up.[31] In the seminal randomized trial of temporal lobectomy versus medical management for patients with drug-resistant temporal lobe epilepsy (defined as monthly seizures for 1 year despite two or more antiseizure medications), 8% of the medical group had no impaired-awareness seizures during the 1-year study period.[32]

Our results highlight this complex and dynamic prognosis of epilepsy for most patients. The majority of our patients had a mix of seizure-free and recent-seizure visits, with highly variable seizure-free intervals between seizures. This challenges the concept that a person's epilepsy is either drug-responsive or drug-resistant, reducing these outcomes to a simple static property. Our findings reflect real-world clinical practice at our center, including factors like prescribed medication changes, patient adherence, and other factors influencing seizure outcomes. We believe that the outcomes observed here are representative of epilepsy treated at academic medical centers, and in line with prior epidemiological studies, arguing that a complex, remitting–relapsing natural history may be more common than is often recognized. Prior studies that addressed similar questions have relied on manual chart reviews, at tremendous effort and cost. NLP can now accomplish these tasks with much lower cost and effort, and at much greater scale. Should our findings bear out at other clinical centers and in other settings, the knowledge that most patients undergo a relapsing–remitting course could greatly affect our approach to epilepsy care. It might also influence epilepsy research to try different approaches to epilepsy treatment in the hope of breaking this cycle.

Patients with well-controlled epilepsy had less frequent and fewer total office visits, but spent more time at our institution than patients who were predominantly having seizures. This finding may seem counterintuitive; one might expect patients who continue to have seizures to be more likely to continue to follow up with their providers. It could be that patients with well-controlled epilepsy return for refills and laboratory tests, whereas patients with poorly controlled epilepsy are dissatisfied with their care and seek care elsewhere; patient satisfaction is positively associated with quality and efficacy of care, and negatively associated with probability of provider change.[33–35]

Our study had several limitations. Our NLP algorithm was developed and tested at a single center. Although our cohort included notes from many unique authors with different writing styles, it will be important for future studies to test generalizability across institutions, and throughout our health system beyond the epilepsy center. Our center is a tertiary academic center, with presumed bias toward more complex and drug-resistant epilepsies. Our method's temporal resolution is limited to clinic visits, rather than individual seizures, so it lacks the temporal granularity of seizure diaries that contain the exact date and time of

every seizure. The observation that seizure-free patients have less frequent and fewer total office visits may bias analysis at the visit level (e.g., inflating the proportion of total visits classified as not seizure-free) but should not bias our analyses at the patient level. Furthermore, as length of follow-up stay is inversely correlated with overall difficulty of epilepsy control, the latter a variable that we can only partially capture, our Kaplan–Meier analysis may not have perfectly noninformative censoring, potential introducing bias into our results. Our findings at the cohort level (e.g., relapse rate of 50% at 3 years and 75% at 10 years) may not apply to all individual patients; selected subgroups may have different prognoses. Our analyses did not account for treatments, including changes in antiseizure medications, which will be an important opportunity for future study. We also identified several opportunities to improve the model's performance, such as accounting for multiple seizure types in a single note, although these limitations did not reduce the model's overall performance below the accuracy of human readers.

5 | CONCLUSIONS

In conclusion, extraction of clinically important seizure outcome measures is feasible using NLP of clinical notes. At our center, the disease course for many patients followed a remitting and relapsing pattern; the majority of patients did not achieve sustained seizure freedom. This method represents a powerful new tool for clinical research with many potential uses and extensions to other clinical questions, including studies for quality assurance, retrospective clinical trials, and rapid patient selection to improve the efficacy of prospective diagnostic and therapeutic investigation.

AUTHOR CONTRIBUTIONS

Kevin Xie, Ryan S. Gallagher, Dan Roth, Brian Litt, and Colin A. Ellis designed the study. Kevin Xie, Russell T. Shinohara, Dan Roth, Brian Litt, and Colin A. Ellis implemented the study. Sharon X. Xie provided statistical expertise for time-to-event analysis. Ryan S. Gallagher, Sharon X. Xie, Chloe E. Hill, Erin C. Conrad, and Kathryn A. Davis provided feedback on the methods, design, and manuscript of the study. All authors were involved in the drafting and editing of the final manuscript.

ACKNOWLEDGMENTS

This research was funded by the National Institutes of Health (NIH) National Institute of Neurological Disorders and Stroke (1DP1 OD029758), by the Mirowski Family Foundation, and by contributions from Jonathan and Bonnie Rothberg. C.A.E. is supported by the NIH National Institute of Neurological Disorders and Stroke (award number K23NS121520), by the American Academy of Neurology Susan S. Spencer Clinical Research Training Scholarship, and by the Mirowski Family Foundation. D.R.'s work was partially funded by the Office of Naval Research (contract N00014-19-1-2620).

CONFLICT OF INTEREST STATEMENT

None of the authors has any conflict of interest to disclose. 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.

DATA AVAILABILITY STATEMENT

Our models are available on the Hugging Face Hub at: https://huggingface.co/CNT-UPenn. Our code is available on GitHub at: https://github.com/penn-cnt/Text_Mining_Epilepsy_Outcomes. We cannot make our clinical data available publicly due to patient privacy.

ORCID

Kevin Xie https://orcid.org/0000-0003-1849-2085
Chloe E. Hill https://orcid.org/0000-0001-5307-4167
Erin C. Conrad https://orcid.org/0000-0001-8910-1817
Colin A. Ellis https://orcid.org/0000-0003-2152-8106

REFERENCES

1. Patel AD, Baca C, Franklin G, Herman ST, Hughes I, Meunier L, et al. Quality improvement in neurology: epilepsy quality measurement set 2017 update. Neurology. 2018;91(18):829–36.
2. Clary HM, Josephson SA, Franklin G, Herman ST, Hopp JL, Hughes I, et al. Seizure frequency process and outcome quality measures: quality improvement in neurology. Neurology. 2022;98(14):583–90.
3. Yew ANJ, Schraagen M, Otte WM, van Diessen E. Transforming epilepsy research: a systematic review on natural language processing applications. Epilepsia. 2023;64(2):292–305.
4. Xie K, Gallagher RS, Conrad EC, Garrick CO, Baldassano SN, Bernabei JM, et al. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. J Am Med Inform Assoc. 2022;29(5):873–81.
5. Xie K, Litt B, Roth D, Ellis CA. Quantifying clinical outcome measures in patients with epilepsy using the electronic health record [Internet]. In: Proceedings of the 21st workshop on biomedical language processing. Dublin, Ireland: Association for Computational Linguistics; 2022 p. 369–375. Available from: https://aclanthology.org/2022.bionlp-1.36
6. Alsentzer E, Murphy J, Boag W, W-H Weng, D Jindi, T Naumann, et al. Publicly available clinical BERT embeddings [Internet]. In: Proceedings of the 2nd clinical natural language processing workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019 p. 72–78. [cited 2022 Aug 25] . Available from: https://aclanthology.org/W19-1909
7. Liu Y, Ott M, Goyal N, J Du, M Joshi, D Chen, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Internet].

2019 [cited 2022 Aug 25] . Available from: http://arxiv.org/abs/1907.11692
8. Vaswani A, Shazeer N, Parmar N, J Uszkoreit, L Jones, AN Gomez, et al. Attention is all you need [Internet]. In: Advances in neural information processing systems. Red Hook, NY: Curran Associates, Inc.; 2017 [cited 2022 Apr 12] . Available from: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
9. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding [Internet]. In: Proceedings of the 2019 conference of the north American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019 p. 4171–4186. [cited 2021 Oct 22] . Available from: https://aclanthology.org/N19-1423
10. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.
11. Battiti R, Colla AM. Democracy in neural nets: voting schemes for classification. Neural Netw. 1994;7(4):691–707.
12. Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst Mag. 2006;6(3):21–45.
13. Fisher RA. Statistical methods for research workers. 12th ed. Edinburgh: Oliver and Boyd; 1954.
14. Liljequist D, Elfving B, Skavberg RK. Intraclass correlation – a discussion and demonstration of basic features. PLoS One. 2019;14(7):e0219854.
15. Deng L, Liu Y. A joint introduction to natural language processing and to deep learning [Internet]. In: Deng L, Liu Y, editors. Deep learning in natural language processing. Singapore: Springer; 2018 p. 1–22. [cited 2023 Feb 9]. Available from: https://doi.org/10.1007/978-981-10-5209-5_1
16. Decker BM, Hill CE, Baldassano SN, Khankhanian P. Can antiepileptic efficacy and epilepsy variables be studied from electronic health records? A Review of Current Approaches. Seizure. 2021;85:138–44.
17. Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open. 2019;9(4):e023232.
18. Decker BM, Turco A, Xu J, Terman SW, Kosaraju N, Jamil A, et al. Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record. Seizure. 2022;101:48–51.
19. Kwan P, Brodie MJ. Early identification of refractory epilepsy. N Engl J Med. 2000;342(5):314–9.
20. Annegers JF, Hauser WA, Elveback LR. Remission of seizures and relapse in patients with epilepsy. Epilepsia. 1979;20(6):729–37.
21. Cockerell OC, Sander JWAS, Hart YM, Shorvon SD, Johnson AL. Remission of epilepsy: results from the National General Practice Study of epilepsy. Lancet. 1995;346(8968):140–4.
22. Sillanpää M, Schmidt D. Natural history of treated childhood-onset epilepsy: prospective, long-term population-based study. Brain. 2006;129(3):617–24.
23. Berg AT, Rychlik K, Levy SR, Testa FM. Complete remission of childhood-onset epilepsy: stability and prediction over two decades. Brain. 2014;137(12):3213–22.
24. Beghi E, Beretta S, Carone D, Zanchi C, Bianchi E, Pirovano M, et al. Prognostic patterns and predictors in epilepsy: a multicentre study (PRO-LONG). J Neurol Neurosurg Psychiatry. 2019;90(11):1276–85.
25. Berg AT, Lin J, Ebrahimi N, Testa FM, Levy SR, Shinnar S. Modeling remission and relapse in pediatric epilepsy: application of a Markov process. Epilepsy Res. 2004;60(1):31–40.
26. Berg AT, Rychlik K. The course of childhood-onset epilepsy over the first two decades: a prospective, longitudinal study. Epilepsia. 2015;56(1):40–8.
27. Bonnett LJ, Powell GA, Tudur Smith C, Marson AG. Breakthrough seizures—further analysis of the standard versus new antiepileptic drugs (SANAD) study. PLoS ONE. 2017;12(12):e0190035.
28. Sillanpää M, Schmidt D, Saarinen MM, Shinnar S. Remission in epilepsy: how long is enough? Epilepsia. 2017;58(5):901–6.
29. Chen H, Amdur R, Pauldurai J, Koubeissi M. Seizure recurrence after prolonged seizure control: patterns and risk factors. Epilepsy Behav. 2021;124:108330.
30. Schiller Y. Seizure relapse and development of drug resistance following long-term seizure remission. Arch Neurol. 2009;66(10):1233–9.
31. Callaghan B, Schlesinger M, Rodemer W, Pollard J, Hesdorffer D, Allen Hauser W, et al. Remission and relapse in a drug-resistant epilepsy population followed prospectively: drug-resistant epilepsy population. Epilepsia. 2011;52(3):619–26.
32. Wiebe S, Blume WT, Girvin JP, Eliasziw M. A randomized, controlled trial of surgery for temporal-lobe epilepsy. N Engl J Med. 2001;345(5):311–8.
33. Glickman SW, Boulding W, Manary M, Staelin R, Roe MT, Wolosin RJ, et al. Patient satisfaction and its relationship with clinical quality and inpatient mortality in acute myocardial infarction. Circ Cardiovasc Qual Outcomes. 2010;3(2):188–95.
34. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open. 2013;3(1):e001570.
35. Marquis MS, Davies AR, Ware JE. Patient satisfaction and change in medical care provider: a longitudinal study. Med Care. 1983;21(8):821–9.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Xie K, Gallagher RS, Shinohara RT, Xie SX, Hill CE, Conrad EC, et al. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. Epilepsia. 2023;64:1900–1909. https://doi.org/10.1111/epi.17633