HHS Public Access

Author manuscript

Epilepsy Res. Author manuscript; available in PMC 2025 November 01.

Published in final edited form as:

Epilepsy Res. 2024 November ; 207: 107451. doi:10.1016/j.eplepsyres.2024.107451.

Extracting Seizure Control Metrics from Clinic Notes of Patients with Epilepsy: a Natural Language Processing Approach

Marta Fernandes, PhD[a,b,*], **Aidan Cardall**[a,b,*], **Lidia MVR Moura, MD, MPH, PhD**[a,b], **Christopher McGraw, MD**[a,b], **Sahar F Zafar, MD**[a,b], **M Brandon Westover**[b,c]

[a]Department of Neurology, Massachusetts General Hospital (MGH), Boston, Massachusetts, United States

[b]Harvard Medical School, Boston, Massachusetts, United States

[c]Beth Israel Deaconess Medical Center (BIDMC), Boston, Massachusetts, United States

Abstract

Objectives: Monitoring seizure control metrics is key to clinical care of patients with epilepsy. Manually abstracting these metrics from unstructured text in electronic health records (EHR) is laborious. We aimed to abstract the date of last seizure and seizure frequency from clinical notes of patients with epilepsy using natural language processing (NLP).

Methods: We extracted seizure control metrics from notes of patients seen in epilepsy clinics from two hospitals in Boston. Extraction was performed with the pretrained model RoBERTa_for_seizureFrequency_QA, for both date of last seizure and seizure frequency, combined with regular expressions. We designed the algorithm to categorize the timing of last seizure ("today", "1 to 6 days ago", "1 to 4 weeks ago", "more than 1 to 3 months ago", "more than 3 to 6 months ago", "more than 6 to 12 months ago", "more than 1 to 2 years ago", "more than 2 years ago") and seizure frequency ("innumerable", "multiple", "daily", "weekly", "monthly", "once per year", "less than once per year"). Our ground truth consisted of structured questionnaires filled out by physicians. Model performance was measured using the areas under the receiving operating characteristic curve (AUROC) and precision recall curve (AUPRC) for categorical labels, and median absolute error (MAE) for ordinal labels, with 95% confidence intervals (CI) estimated via bootstrapping.

Results: Our cohort included 1,773 adult patients with a total of 5,658 visits with reported seizure control metrics, seen in epilepsy clinics between December 2018 and May 2022. The cohort average age was 42 years old, the majority were female (57%), White (81%) and non-Hispanic (85%). The models achieved an MAE (95% CI) for date of last seizure of 4 (4.00–4.86) weeks, and for seizure frequency of 0.02 (0.02–0.02) seizures per day.

---

*Co-first authors

Declarations of interest

None.

Conclusions: Our NLP approach demonstrates that the extraction of seizure control metrics from EHR is feasible allowing for large-scale EHR research.

Keywords

Epilepsy; Phenotyping; Seizure control; Electronic health records; Natural language processing; Large language model

1. Introduction

Time since last seizure occurred and frequency of seizures are important metrics for providers to guide clinical care of patients with epilepsy (Patel et al., 2018; Munger Clary et al., 2022). To extract these metrics automatically from electronic health records (EHR) clinical notes reliably for research and quality surveillance purposes is still a challenge because it entails laborious manual chart review. The unstructured nature of clinical notes, coupled with medical history of multiple types of seizures and their corresponding seizure frequencies, medical abbreviations, and other text properties (Reiter & Dale, 1997; Pevy et al., 2021) makes extraction of these seizure control metrics even more challenging.

Natural language processing (NLP) aims to teach machines how to read human language from structured or unstructured text efficiently, enabling expansion of clinical research through large-scale automated chart review. Recent NLP work (Xie, Gallagher, et al., 2022; Xie, Litt, et al., 2022) introduced pretrained language models, based on the Transformer architecture (Vaswani et al., 2017), to extract seizure frequency and date of most recent seizure from the text of outpatient progress notes for patients with epilepsy. Models were pretrained on a large general text corpus to obtain a general understanding of language and then finetuned on small datasets to adapt them for specific tasks or domains (Han et al., 2021). The authors (Xie, Gallagher, et al., 2022; Xie, Litt, et al., 2022) used clinical notes to fine-tune Transformer models, achieving high accuracy at extracting sections of text which contained the answers to three key questions: whether a patient was seizure free at the time of a given clinic visit (median accuracy 84%); and what was their seizure frequency (accuracy 88%, F_1 score 85%) and date of most recent seizure (accuracy 86%, F_1 score 83%). However, this work had important limitations: First, the ‘ground truth’ was based on chart review rather than an independent source of truth. Second, evaluation was limited to a single institution. In the present work, we build upon this prior work by addressing these limitations.

In our study, we used the aforementioned fine-tuned model to extract date of most recent seizure and seizure frequency. We extract text containing the answer, and also provide additional processing (regular expressions) that yields explicit, structured answers, suitable for large-scale EHR research studies. Second, for the gold standard we utilize direct questionnaires (independent of the unstructured note) where doctors explicitly answered the question of interest, thus providing an improved gold standard. Lastly, our study represents a test of the previous author’s model at a different institution and is thus a test of how well the model generalizes to a dataset from a different hospital.

Our main goal in this study is to extend and externally validate pretrained and finetuned NLP models to abstract the date of last seizure and seizure frequency from unstructured clinical notes of patients with epilepsy. This methodology has the potential to allow for rapid large-scale retrospective research through automatic extraction of seizure control metrics from unstructured clinical notes.

2. Methods

2.1. Study cohort

This study is reported in accordance with the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) statement (Vandenbroucke et al., 2014). EHR data was extracted under a protocol approved by the Mass General Brigham (MGB) Institutional Review Board with a waiver of informed consent. In the study, we conducted a retrospective analysis of clinic notes from adult patients (≥18 years old) with a diagnosis of epilepsy seen in epilepsy clinics at Mass General Hospital (N = 4,013) and Brigham and Women's hospital (N = 1,645) between December 28th 2018 and May 20th 2022. The clinic notes for each visit included all notes for the visit date, such as progress notes, consult notes and hospital course. The ground truth for the diagnosis of epilepsy was extracted from structured questionnaires filled out by physicians (Fernandes et al., 2022), as part of a quality improvement project embedded in standard of care. These questionnaires contained fields where the physician indicated, among other seizure control metrics, the date of last seizure, seizure frequency, and seizure types (Donahue et al., 2021). Up to four seizure types could be recorded with their corresponding frequencies and dates of last seizure (Donahue et al., 2021). The seizure control metric information contained in these questionnaires was considered our ground truth.

2.2. Inclusion and exclusion criteria

Candidates for our cohort included all patients with epilepsy questionnaires filled out regarding seizure control metrics. Based on our inclusion and exclusion criteria (Figure 1), we removed all cases in which the questionnaires did not provide at least one date of last seizure or one seizure frequency. We then created a dataset for each seizure control metric, where again each note had to provide at least a value for the metric.

2.3. Modeling approaches

Our objective was to extract from unstructured notes patients' date of last seizure and seizure frequency. We utilized two approaches (Table 1): *Semi-direct:* using a regex to select a portion of text for a given "prompt" followed by a regular expression (regex) to map the text to a final value of the seizure control metric of interest; *Indirect*: Transformers followed by regular expressions (i.e. using a Transformer model to select relevant text, then a regex to map the text to a final value of the seizure control metric of interest). Timing of the last seizure and seizure frequency were extracted using either the semi-direct or indirect approaches. The regexes and Transformer prompts used for each task are shown in Table 1. The corresponding labels are shown in Table 2 (Fernandes et al., 2022). The main regexes used to map the text to a final value of the seizure control metric of interest are presented in Table A.1. The regexes were designed in a flexible way, so that both

quantitative and non-quantitative values could be captured and assigned the respective label. Thus, if the expression was “few days ago”, the category assigned for date of last seizure was “1 day to 6 days ago” (‘1DAY’). We also designed regexes to capture dates written in different ways, which we include in our code publicly available in a GitHub repository. Notes preprocessing consisted of removal of special characters and blank spaces, followed by lowercasing (Fernandes et al., 2023).

Semi-direct approach: In the semi-direct approach, for each prompt, we captured the corresponding portion of text following the expression with a total length of 50 tokens (words). For example, for the prompt “Last seizure”, the regex captured “last seizure was over 2 years ago and she has remained seizure free”. We then applied a function with a series of regexes to map the captured text to a value of the seizure control metric (Table 2). For this given example, the label “2YR” (more than two years ago) was assigned.

Indirect approach: In the indirect approach, we used RoBERTa_for_seizureFrequency_QA (Xie, Gallagher, et al., 2022), which is a finetuned model for extraction of both date of last seizure and seizure frequency. This model is based on the original RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019) pretrained model, which is in turn an enhanced version of BERT with improved pretraining objectives and hyperparameters. In this work, we used the hyperparameters from prior work (Xie, Gallagher, et al., 2022) without any further hyperparameter optimization. The “prompts” in Table 1 with a question mark, such as “When was the patient last seizure?” were prompts to the Transformer.

With each prompt, the Transformer extracted for each note a small amount of text that (likely) contained the information of interest. The text extracted by the Transformer (the “answer” text) was subsequently mapped (where possible) to a final label shown in Table 2, as in the semi-direct approach.

We combined the answers from both approaches for evaluation of modeling performance, thus we refer to model outputs as those resulting from the indirect or semi-direct approaches.

For date of last seizure, in case the model answer was an actual calendar date, instead of selecting one of the coded options (e.g. 1WK – one to four weeks ago, or 6MON – more than six to twelve months ago), the regex captured the date, subtracted it from the date of the clinic visit, and assigned the closest matching label. This approach was also applied for model answers regarding seizure frequency. If no label was assigned based on regexes for all other frequencies or dates, in case the regex matching indicated that a patient was seizure free (i.e. the term “seizure free” or “sz free” was detected), we assigned the label “2YR” (more than two years ago) for date of last seizure and “YEAR” (less than once per year) for seizure frequency.

2.4. Modeling performance evaluation

We assessed model performance in two ways: treating the labels as categorical, and as ordinal.

Categorical performance—We assessed model performance with respect to categorical labels for both seizure control metrics. We also assessed models' performance when merging labels based on their proximity. We created four categories for date of last seizure ("TOD-1DAY" – today to six days ago, "1WK-5WK" – one week to three months ago, "13WK-6MON" – more than three to twelve months ago, "1YR-2YR" – more than one year ago). We also created four categories for seizure frequency ("INN-MULT-DAIL" – innumerable, multiple or daily, "WKLY" – weekly, "MNTH" – monthly, "NEM-YEAR" – at least once or less than once per year).

Model performance relative to categorical labels was evaluated using area under the precision-recall curve (AUPRC) and area under the receiver operating characteristic curve (AUROC) (Saito & Rehmsmeier, 2015). AUPRC quantifies the trade-off between precision (also called positive predictive value) and recall (also called sensitivity) for different thresholds. AUROC quantifies the tradeoff between sensitivity and false positive rate (1 – specificity) (Steyerberg et al., 2010). Because of class imbalance, we evaluated the micro average performance, which aggregates the contributions of all classes to compute the average metric. Micro average performance is suited for problems with class imbalance and consists of calculating metrics globally by counting the total true positives, false negatives and false positives.

Physicians could record up to four seizure types with their corresponding frequencies and dates of last seizure for a patient visit. Thus, to evaluate the different model outputs, we assessed the true positive and true negative classifications with respect to each label. For each label we assigned either a 1 (present) or 0 (absent or missing).

Ordinal performance—To assess model performance on an ordinal scale, we assigned numbers considering time in weeks for date of last seizure and number of seizures per day for seizure frequency. Thus, for date of last seizure, we assigned the following numbers to the labels ['TOD', '1DAY', '1WK', '5WK', '13WK', '6MON', '1YR', '2YR']: [0, 1/7, 1, 5, 13, 24, 52, 104]. For seizure frequency, we assigned the following numbers to the labels ['INN', 'MULT', 'DAIL', 'WKLY', 'MNTH', 'NEM', 'YEAR']: [2, 4×2/7, 4/7, 12×4/365, 12/365, 10/365, 1/365]. For evaluation of model performance with respect to these ordinal labels, we created a scatter plot with true vs assigned labels, and assessed the distribution of the absolute errors ($|y_{true} - y_{assigned}|$) between the models ($y_{assigned}$) and ground truth ($y_{true}$) ordinal labels.

Since one up to four seizure frequencies and dates of last seizure could be recorded by the physician, we evaluated each individual model output against the closest ground truth, for ordinal labels. Thus, for each model output, we selected the closest ground truth, in case there was more than one, and computed the corresponding absolute error for each model ordinal label. As an example, for a model output with date of last seizure '1DAY', and ground truth '1DAY' and '1WK', the error for this model would be 0, since this label is included in the ground truth. If we changed the ground truth to only '1WK', the error would be 1 − 1/7~0.86, since we assigned the values 1 to '1WK' and 1/7 to '1DAY'. For a model output with no label (absent or missing) we assigned the closest error, e.g. for ground truth

‘1DAY’ the error would be 1/7, and for ground truth ‘1DAY’ and ‘1WK’, the error would be 1/7 (the closest).

Confidence intervals—We calculated 95% confidence intervals for all performance metrics via 1000 bootstrapping iterations.

3. Results

3.1. Study cohort characteristics

The study cohort included 1,773 patients with epilepsy (Table 3). Average age of 42 years old. The majority were females (57%), White (81%) and non-Hispanic (85%). After applying inclusion and exclusion criteria (Figure 1), we observed that seizure frequency included a smaller number of patients compared to the other metric, mainly due to several visits reporting that the seizure frequency was not well defined.

Statistics of models answers and extraction regexes for date of last seizure and seizure frequency are presented in Table A.2. We also present in Table A.3. examples of cases where Transformer outputs could not be translated into seizure control metric labels, due to either an ambiguous answer (e.g. related to patient uncertainty “does not know for sure” or “increased seizure frequency”) or lack of sufficient information related to the metric (e.g. “seizure” or “occasional”).

3.2. Evaluation of modeling performance

The full list of modeling results is presented in the appendix (Table A.4. and Figures A.1. and A.2.), where the answers from both semi-direct and indirect approaches were combined for evaluation of modeling performance.

Categorical performance—We evaluated models for date of last seizure and seizure frequency with categorical labels, both individual and merged, as shown in Figure 2. With individual labels, we obtained a micro average AUROC (95% CI) of 0.63 (0.62–0.63) and 0.61 (0.61–0.62), AUPRC (95% CI) of 0.44 (0.43–0.45) and 0.43 (0.42–0.44) for last seizure and seizure frequency, respectively. When we merged labels into smaller groups, AUROC remained approximately the same, but we observed an increase in AUPRC of 16% and 11%, for last seizure and seizure frequency, respectively. Overall, the performance was more balanced after merging the labels.

Ordinal performance—We assessed model performance for ordinal labels, presented in Figure 3. The model achieved a median absolute error (95% CI) of 4 (4.0–4.86) weeks and 0.02 (0.02–0.02) seizures per day, for last seizure and seizure frequency, respectively.

Qualitative error analysis—We performed manual review (MBW) of 10 cases where the RoBERTa models did not match the ground truth date of last seizure. We observed two scenarios for the model errors: in one scenario (60% of cases) there was lack of any clearly documented information in the note; in the other (40% of cases) the model appeared correct based on information in the note, but the note contradicted the information provided in the physician structured questionnaire.

4. Discussion

4.1. Principal findings

We have developed NLP algorithms to abstract date of last seizure and seizure frequency from unstructured clinical EHR notes. Our work builds on a previously published pretrained and finetuned Transformer model and adds to this regular expressions to extract date of last seizure and seizure frequency. We introduced additional processing to transform the free text output into structured data, or labels. A novel aspect of our work is that we were able to evaluate the model output relative to an independent ground truth, based on direct answers from medical doctors to questionnaires that were filled out separately from the unstructured clinical notes. Our work also provides an independent validation in a novel dataset from a different institution, further clarifying the strengths and challenges of using NLP approaches in unstructured clinical notes to extract seizure control metrics.

4.2. Comparison with prior work

The study (Xie, Gallagher, et al., 2022) which developed the finetuned model that we leveraged in our study extracted text containing seizure frequency and date of last seizure from clinical notes. Their finetuned model $\text{RoBERTa}_{\text{FT}}$ achieved good performance when extracting seizure frequency and date of last seizure, with overall F1 scores of 0.85 and 0.83. However, the authors measured text span overlap rather than the prediction's correctness, thus we cannot directly compare that study performance with ours. We believe there are three factors which likely explain our low F1 scores for seizure frequency and date of last seizure (0.37 and 0.38). First, we used the pre-fine-tuned model from the prior study, as a test of generalizability; further fine tuning the model to data from our institution might improve performance. Second, while (Xie, Gallagher, et al., 2022) used progress notes, we used epilepsy clinics notes and any other types of notes (Fernandes et al., 2023); the authors (Xie, Gallagher, et al., 2022) extracted text from progress notes as "paragraphs" to improve the probability that the text involved epileptic events, and truncated these paragraphs to better fit within the maximum sequence length of the models (512 tokens), while our notes had paragraphs but these were not pre-selected or truncated. Third, the prior study evaluated whether the model correctly captured information in notes relative to a human reading of the same notes, whereas we evaluated our models relative to an external gold standard (i.e. questionnaires filled out separately from the notes), which is one step removed.

Other studies (Decker et al., 2022; Fonferko-Shadrach et al., 2019) have developed rule-based NLP algorithms to abstract seizure types and frequencies, among other seizure control metrics, from EHR unstructured clinic notes. In (Decker et al., 2022), the algorithm achieved an overall 22% recall, 73% precision, and 0.40 F1 score to evaluate seizure types and frequencies. In (Fonferko-Shadrach et al., 2019), the model achieved (precision, recall) for seizure frequency (86.3%, 53.6%). The study sample sizes were smaller than our study, with 150/219/248 notes for training/validation/test in (Decker et al., 2022) and 200 clinical letters for validation with 1925 items of information in (Fonferko-Shadrach et al., 2019). Direct comparison between our study and others is limited due to the different methodologies used to define cohorts and the methods of measuring performance. This highlights the need to establish benchmark datasets against which to evaluate models.

4.3. Limitations

Our study has limitations. This study was limited to a single hospital system, in one geographic region (Boston, Massachusetts). Thus, the cohort may not be representative of other US and non-US populations. Another limitation was that some Transformer outputs were not able to be mapped to a label for seizure frequency or date of last seizure. As mentioned above, we used a pre-fine-tuned model from another institution; further fine-tuning could potentially improve performance. However, we also observed that often there were discrepancies between what was indicated in the structured survey, our ground truth, and the unstructured clinical note. Therefore, to more fairly evaluate the model, each case would have to be manually reviewed to reduce “label noise”, which is unfortunately infeasible at scale. Nevertheless, our analysis using ordinal labels suggests that the error between ground truth and models’ answers was generally small and acceptable. Future studies should focus on developing larger, more highly cleaned test sets and testing generalizability and improving models to work robustly across hospitals.

4.4. Conclusions

Our NLP approach extracted seizure control metrics including date of last seizure and seizure frequency with reasonable accuracy from unstructured clinical notes of patients with epilepsy. This methodology can enable large-scale research that relies on seizure control metrics extracted from EHR data.

Acknowledgements

Dr. Westover’s laboratory is supported by grants from the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598) and NSF (2014431). Dr. Sahar F. Zafar is a clinical neurophysiologist for Corticare, unrelated to this work and was supported by the NIH (K23NS114201). Dr. Lidia M. V. R. Moura was supported by the NIH (NIH-NIA 5K08AG053380-02, NIH-NIA 5R01AG062282-02, NIH-NIA 2P01AG032952-11, NIH-NIA 3R01AG062282-03S1, NIH-NIA 1R01AG073410-01), and the Epilepsy Foundation of America. We acknowledge the Epilepsy Learning Health System (ELHS).

Appendix

Table A.1.

Regular expressions for seizure control metrics extraction after removal of special characters from the notes.

Label	Regular expression
Date of last seizure	
Today	• ‘(was)?\s?today’ • ‘this morning’ • ‘this am’ • ‘nightly’
1 day to 6 days ago	• ‘1–6\s?day’ • ‘\d{3,4} [a|p]m’ • ‘tues wed thur fri sat sun’ • ‘daily’ • ‘(was)?\s?yesterday’ • ‘most recent (\w+\s)+\w+day’ • ‘1–6\s?d’ • ‘.*?\s?[1–5]{1}-[2–6]{1}\s?d’

Label	Regular expression
	• '.*?\s?(([1–7]{1})|(one)|(two)|(three)|(four)|(five)|(six))\s?d(ay)?(s?)' • 'this week' • 'few days ago' • 'weekly' • 'day of the visit' • '.*?\s?(a|1|(one))\s?week' • '.*?\s?(hour(s)?|night(s)?)' • 'last \w+day' • 'twice a day'
1week to 4 weeks ago	• 'more than (a|1|(one)) week to (2|3|4|(two)|(three)|(four)) weeks' • '.*?\s?last week' • 'weekend' • 'this past sun' • 'last week' • 'last was week' • 'past week' • 'couple (of)?\s?weeks' • 'few weeks' • '.*?\s?(([1–4]{1})|a|(one)|(two)|(three)|(four))\s?w(ee)?k(s?)' • '.*?\s?[1–3]{1}-[2–4]{1}\s?w(ee)?k(s?)' • '.*?\s?(a|1|(one))\s?mo(nth)?' • '.*?\s?([7–9]|[1][0–9]|[2][0–9])\s?d(ay)?s?' • 'last month' • 'past month' • '>?\s?[1–4]\s?week' • 'this month' • 'each month' • 'end of the month' • 'weeks ago' # last episode weeks ago
More than 1 month to 3 months ago	• 'more than 1 month to 3 months' • '5 weeks' • '.*?\s?5–12\s?w' • '.*?\s?[4–9]-[5–9]\s?w(ee)?k(s?)' • '.*?\s?[4–9]-[1][0–2]\s?w(ee)?k(s?)' • '.*?\s?(([5–9]{1}|[1][0–2])|(five)|(six)|(seven)|(eight)|(nine)|(ten)|(eleven)|(twelve))\s?w(ee)?k(s?) ago' • '.*?\s?[1–2]-[2–3] mo(nths)?\s?ago' • 'few?\s? mo(nths)?\s?ago' • 'last few?\s? mo(nths)?' • 'couple of months ago' • 'couple (of)?\s?months' • '.*?\s?(([1–3]{1})|(one)|(two)|(three))\s?mo(nths)?' • '>\s?[1–3]\s?m'
More than 3 months to 6 months ago	• 'more than 3 mo(nths)? to 6 mo(nths)?' • 'several weeks ago' • '.*?\s?13–26\s?w' • '.*?\s?[3–5]-[4–6]\s?mo(nths)?' • '.*?\s?(([4–6]{1})|(four)|(five)|(six))\s?mo(nths)?' • '>\s?[3–5]\s?m'
More than 6 months to 12 months ago	• 'more than 6 months to 12 months' • 'several months ago' • '.*?\s?6–12\s?mon' • '.*?\s?[7–9]-[8–9]\s?mo(nths)?' • '.*?\s?[7–9]-[1][0–2]\s?mo(nths)?' • '.*?\s?([7–9]|[1][0–2])\s?mo(nths)?' • '.*?\s?((one)|1)\s?y(ea)?r?' • '>\s?[6–9]\s?m' • '>\s?[1][0–2]\s?m'
More than 1 year to 2 years ago	• 'more than (a|1)\s?y(ea)?r' • '.*?\s?13–24\s?mon' • 'less than ((two)|2|a)\s?y(ea)?r?(s)? ago' • '.*?\s?\>?\s?(1|a|(one))\s?y(ea)?r?' • 'last (1[3–9])|(2[0–3]) months' • '>\s?1\s?/?y' • 'last year' • 'past year'
More than 2 years ago	• 'more than 2 y(ea)?r(s)?' • '.*?\s?([3–9]|[1][0–9]|[2][0–9]|[3][0–9]|[4][0–9])(\+?)\s?y(ea)?r?(s)?' • '.*?\s?(([3–9]|[1][0–9]|[2][0–9]|[3][0–9]|[4][0–9])-([3–9]|[1][0–9]|[2][0–9]|[3][0–9]|[4][0–9]))(\

Label	Regular expression
	+?)\s?y(ea)?r?(s)?' • '>\s?2\s?/?y' • 'in years' • 'age \d{1,2}' • '(was)?\s?((a few)|(several)|(many))?\s?yfea)?r?fs)?'
Seizure frequency	
Innumerable	• 'innumerable' • '50100 (seizures)?\s?per day' • '2030 (seizures)?(events)?\s?per day' • 'high (seizure)?\s?frequency' • '23 per day'
Multiple per day	• 'multiple' • 'times a?(per)? day' • 'twice a day' • '6 per day' • '[2–9]a?\s?day' • '[2–9] nightly' • '[2–9]\s?(events)?(seizures)? a day' • 'five seizures day' • '56 spells per day' • 'up to 2 every 8 hours' • 'several in a day'
Daily	• 'daily' • 'once a day' • 'every day' • 'every morning' • 'every evening' • 'every night' • 'in (one)?1? day' • 'spells a day' • '1 (seizure)?(per)?a?\s?day'
Weekly but not daily	• 'weekly' • 'weekly but not daily' • '([\w\d]).{1,15}\s?week' • 'every [3–9] days' • 'every 810 days' • '[1–3] nocturnal seizures?' • 'events on tues wed thur fri sat sun' • 'every few days' • 'one to a few seizures every couple of nights' • '79 events in 30 days'
Monthly but not weekly	• 'at least once per month' • 'monthly but not weekly' • '([\w\d]).{1,15}\s?month' • 'month' • '13 times each year' • '12xyear' • '34 per year' • 'currently 1 q 5 wks' • 'freq 1 q 2wk'
At least once per year, but not every month	• 'at least once per y' • 'per year' • 'a year' • 'every year' • '([1–9])?(seizures)?(events)?\s?in the past year' • '\d\/\s?year' • '<\s?1\/\s?/?y' • 'yearly' • 'in (the)?\s?last year' • 'one seizure q 60 days' • 'over the past year' • 'couple of (times)?\s?a?(per)?\s?year' • 'less than 1 q 3mos' • 'a few times year' • 'one to two events each year' • '2 seizures over the course of the last year' • '6 to 8 total seizures in three years'

Label	Regular expression
	• ‘last year she has had 5 small seizures’ • ‘2 gtc year’
Less than once per year	• ‘[1–9] y(ea)?r’ • ‘>\s?2\s?/?y’ • ‘sz\s?free’ • ‘seizure\s?free’ • ‘event\s?free’ • ‘for years’ • ‘no (sz?)?(seizure?)?(s?)? since \d\d\d\d’ • ‘less than 2 years ago’ • ‘none for 2?\s?years’ • ‘none recently’ • ‘2year seizure recurrence’ • ‘none in (many)?(several)?\s?years’ • ‘less than onc?e per y’ • ‘onc?e (or)?(twice)?\s?a?\s?y’ • ‘every (other)?(few)? year’ • ‘in (his)?(her)?\s?life’ • ‘life\s?time’ • ‘three in adult life’ • ‘frequency none’ • ‘since (19)?(20)?’ • ‘frequency is approximately 1.53’ • ‘once in a blue moon’ • ‘at least one seizure each year’ • ‘over his life he has had a total of 3’ • ‘no seizures for many years’ • ‘rarely’ • ‘no (gtc)?(seizures)?\s?in years’ • ‘seizure free’ • ‘sz free’ • ‘seizurefree’ • ‘event free’

Table A.2.

Statistics of models answers and extraction regexes for date of last seizure and seizure frequency.

Question metric	No. extractions (%)	No. labels assigned (%)	Extraction method
Date of last seizure			
Last_seizure	5582 (100)	5196 (93.1)	RoBERTa – *indirect approach*
Recent_seizure	5582 (100)	5185 (92.9)	
Last_event	5582 (100)	5204 (93.2)	
Recent_event	5582 (100)	5197 (93.1)	
Last seizure	1647 (29.5)	1515 (27.1)	Regexes – *semi-direct approach*
Last sz	14 (0.3)	14 (0.3)	
Last convulsion	29 (0.5)	27 (0.5)	
Last event	717 (12.8)	707 (12.7)	
Seizure frequency			
Often_event	4732 (100)	3764 (79.5)	RoBERTa – *indirect approach*
Often_seizure	4732 (100)	3835 (81.0)	
Frequency_seizure	4732 (100)	3617 (76.4)	
Frequency_event	4732 (100)	3644 (77.0)	

Question metric	No. extractions (%)	No. labels assigned (%)	Extraction method
Seizure frequency	1126 (23.8)	655 (13.8)	Regexes – *semi-direct approach*

Table A.3.

Examples of cases from the Transformer outputs where seizure control metrics could not be translated into a label.

Last_seizure **answer**	***Seizure_frequency*** **answer**
Indication of event/frequency	
he had another seizure	seizures still cycle
her most recent event	relatively regular rate of seizures
her most recent seizure	every now and then
continuing to have seizures	every other day
he did have another episode	average frequency of every other night
she did have an odd event	freq. every other night
he did have another episode	all the time
	more frequent when waking up
	most often in the afternoon
Indication of no event/frequency	
he has had no spells since	none since
has had none since	no further seizures.
none since	
none for that period of time	
not had a seizure in some time	
Reference to last visit	
5 seizures since last visit	the same frequency
has had 3 seizures since the last visit	about the same rate as before
no additional events since our last encounter	
no events since last seen	
no seizures since last clinic visit	
since last visit has had 2 starring spells	
two events since last seen	
two seizures since last encounter	
Number and types of events	
1 gtc since in the week after	1 gtc
1 possible brief seizure	absence seizures every now and then
	1frequent left temporal epileptiform
2 tonicclonic seizures	discharges 2
	facial spasm was quite frequent
2. freezes	2 other times she will get an aura
34 gtc szmonth	dejavu episodes once in a while
4 gtc.	one aura
complex partial seizures	brief 2 to 3second events of unawareness
one breakout seizure	more frequently he has events of speech arrest
one generalized convulsion	single clear spell
one seizure	frequency 3 seizures in total
partial seizure	she had at least 2 more
	none frequency 3 episodes in total
	3 seizures with 2 3min long and another one a little over 2min
	single seizure
	3 seizures over the course of a couple of days
	2 seizures this year
Medications dosages and frequency	
keppra would be the choice.	nightly
lorazepam ativan 1 mg tablet	nightly at bedtime

***Last_seizure* answer**	***Seizure_frequency* answer**
medication instructions	take 1 tablet by mouth nightly
medication titration schedule	three tablets75mg nightly
	mg tablet take 1 tablet 5 mg total by mouth nightly
	1 tablet by mouth every 4 four hours
	1 tablet by mouth every other day
	2 tablets by mouth every 6 hours if needed
	every 4 four hours as needed
	every 8 eight hours
	seizure activity max 3 mg per day
	1.5 in the am and 2 at night
	once for 1 dose
	once in the morningonce at night
	every 6 six hours
	3 aeds.
	drugresistant epilepsy
Single words	
awareness	fatigued
below	migraines
chair	occasional
coat	blood
compete	awake
oil	days
head	home
himself	coffee
his ketone levels	freedom
plan	disability
practice	recording
seizure	seizures.
not	mgday
Punctuation and single numbers	
.	
.....	.....
2	2
e	3
Other undefined	
since his last visit	seizure frequency
at the last visit	every second or 2 for a while then stops
Patient uncertainty	***Increase or decrease in frequency***
does not know for sure	increased frequency of all 3 types of seizures
she does not recall the last definite event	the frequency and duration of her seizures
concerning for seizure	have increased recently
cant remember the last time he had one	increased seizure frequency
	less frequently
	frequency is now pretty low
	she has still been having startle events much improved in frequency
	he is now having these episodes more frequently
	as he aged they became more frequent

Table A.4.

Micro average performance for categorical labels, performed for 1000 bootstrapping iterations to calculate 95% confidence intervals.

Metric/labels	**AUROC**	**AUPRC**	**Accuracy**	**Specificity**	**Recall**	**F1 score**
Last seizure						

Metric/labels	AUROC	AUPRC	Accuracy	Specificity	Recall	F1 score
Individual	0.63 [0.62–0.63]	0.44 [0.43–0.45]	0.32 [0.31–0.34]	0.90 [0.90–0.91]	0.37 [0.36–0.38]	0.38 [0.37–0.39]
Merged	0.64 [0.64–0.65]	0.60 [0.59–0.61]	0.44 [0.43–0.45]	0.81 [0.81–0.82]	0.50 [0.49–0.51]	0.52 [0.51–0.53]
Seizure frequency						
Individual	0.61 [0.61–0.62]	0.43 [0.42–0.44]	0.28 [0.27–0.30]	0.88 [0.88–0.88]	0.36 [0.35–0.37]	0.37 [0.36–0.38]
Merged	0.61 [0.61–0.62]	0.54 [0.54–0.55]	0.35 [0.34–0.37]	0.78 [0.78–0.79]	0.46 [0.45–0.47]	0.46 [0.45–0.47]

Legend: AUROC – Area under the receiver operating curve; AUPRC – Area under the precision-recall curve.

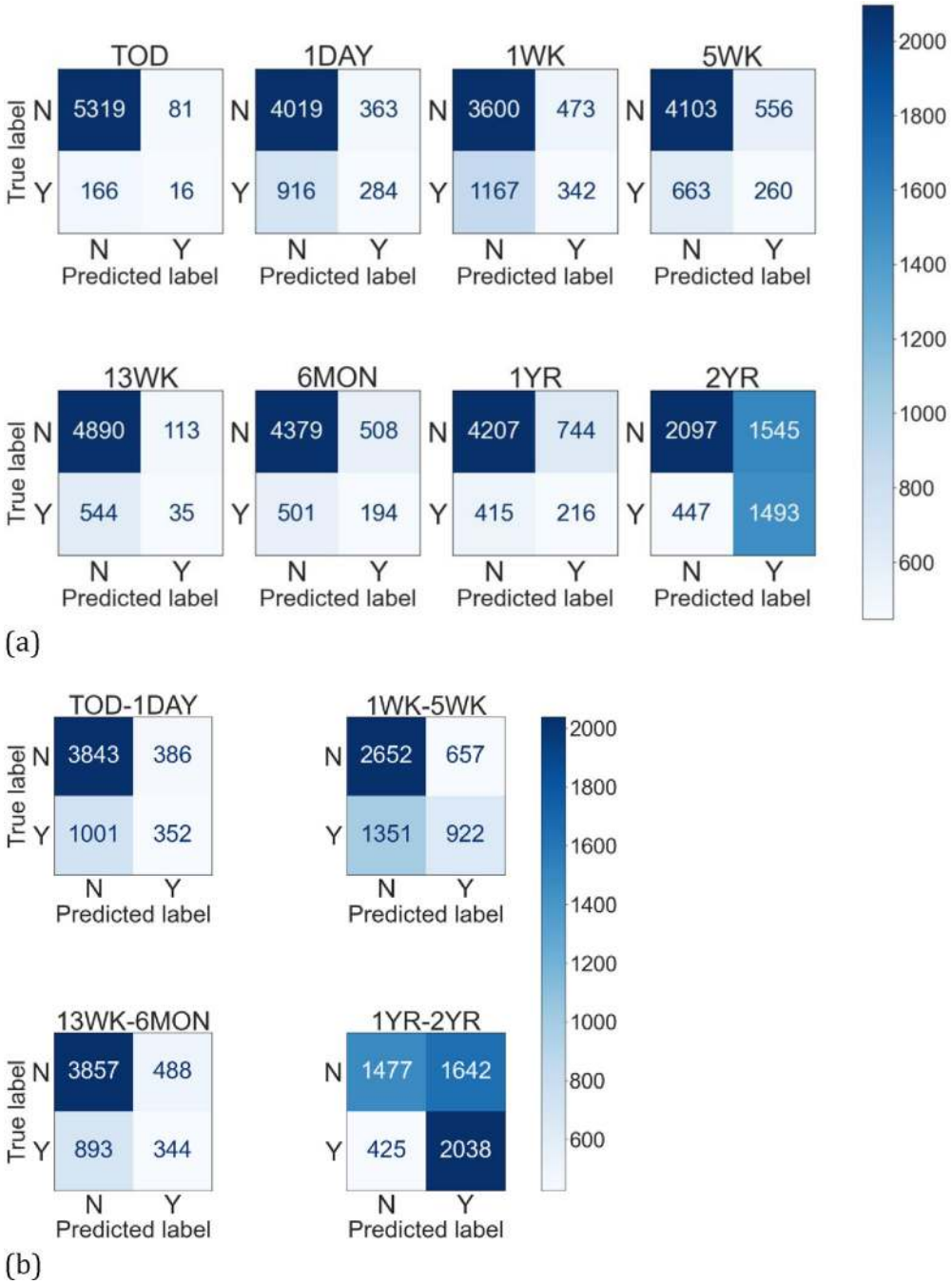

Figure A.1.

Confusion matrices for classification of date of last seizure, with (a) individual and (b) merged labels. Y – yes, N – no. Classifications: YY – true positive, NN – true negative, $Y_{predicted}N_{true}$ − false positive, $Y_{true}N_{predicted}$ − false negative.

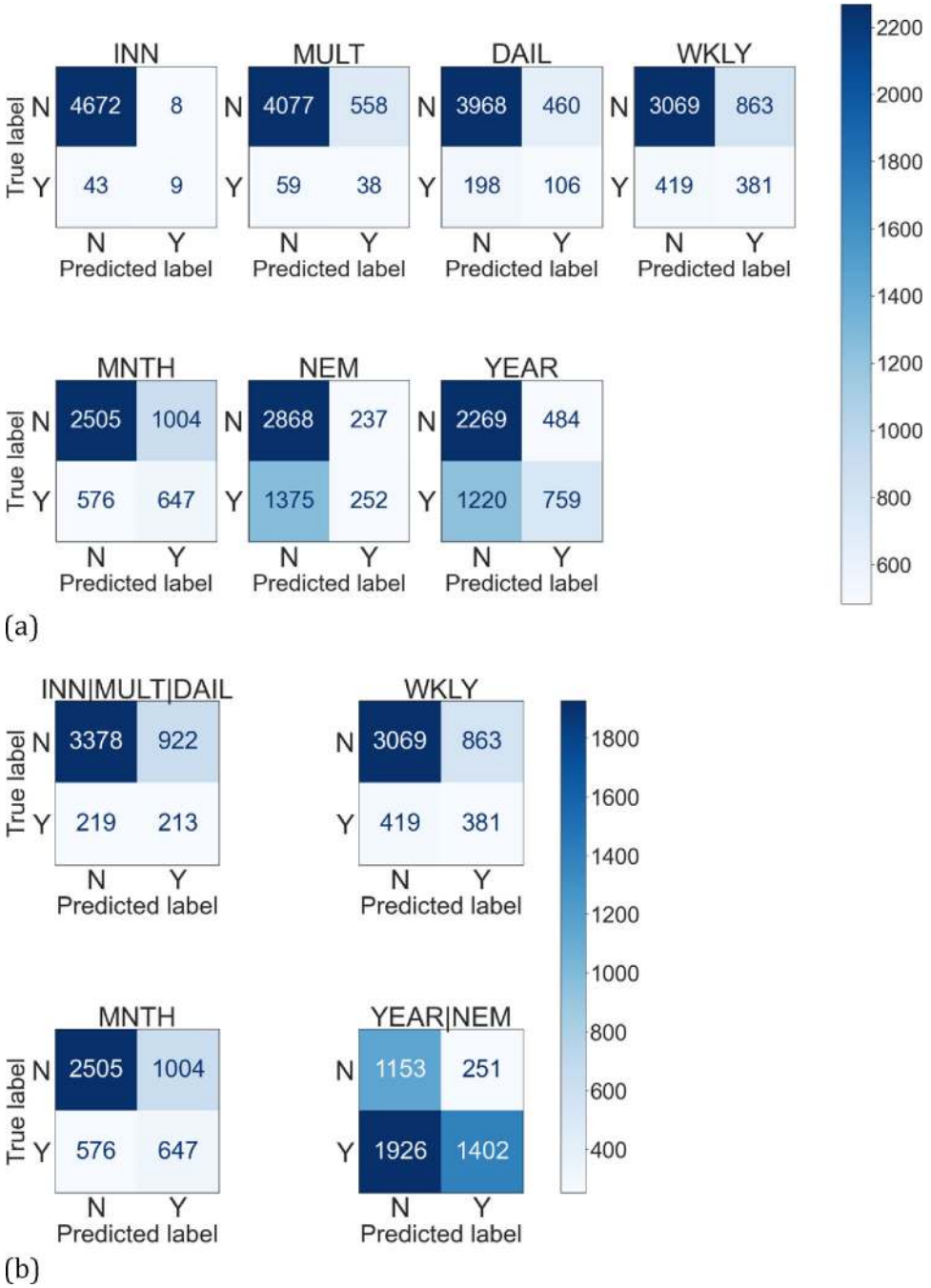

Figure A.2.

Confusion matrices for classification of seizure frequency, with (a) individual and (b) merged labels. Y – yes, N – no. Classifications: YY – true positive, NN – true negative, $Y_{predicted}N_{true}$ – false positive, $Y_{true}N_{predicted}$ – false negative.

Abbreviations:

AUPRC

Area under the precision recall-curve

AUROC
Area under the receiver operating characteristic

CI
Confidence interval

EHR
Electronic health record

MAE
Median absolute error

MGB
Mass General Brigham

NLP
Natural language processing

RoBERTa
Robustly Optimized BERT Pretraining Approach

SD
Standard deviation

STROBE
STrengthening the Reporting of OBservational studies in Epidemiology

Last seizure: TOD
Today

1DAY
One day to six days ago

1WK
One to four weeks ago

5WK
More than one month to three months ago

13WK
More than three months to six months ago

6MON
More than six to twelve months ago

1YR
More than one year to two years ago

2YR
More than two years ago

Seizure frequency: INN
Innumerable

MULT
Multiple per day

DAIL
Daily

WKLY
Weekly

MNTH
Monthly

NEM
At least once per year

YEAR
Less than once per year

References

Decker BM, Turco A, Xu J, Terman SW, Kosaraju N, Jamil A, Davis KA, Litt B, Ellis CA, Khankhanian P, & Hill CE (2022). Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record. Seizure, 101, 48–51. 10.1016/j.seizure.2022.07.010 [PubMed: 35882104]

Donahue MA, Herman ST, Dass D, Farrell K, Kukla A, Abend NS, Moura LMVR, Buchhalter JR, & Fureman BE (2021). Establishing a learning healthcare system to improve health outcomes for people with epilepsy. Epilepsy & Behavior: E&B, 117, 107805. 10.1016/j.yebeh.2021.107805

Fernandes M, Cardall A, Jing J, Ge W, Moura LMVR, Jacobs C, McGraw C, Zafar SF & Westover MB (2023). Identification of patients with epilepsy using automated electronic health records phenotyping. Epilepsia, 64(6), 1472–1481. [PubMed: 36934317]

Fernandes M, Donahue MA, Hoch D, Cash S, Zafar S, Jacobs C, Hosford M, Voinescu PE, Fureman B, Buchhalter J, McGraw CM, Westover MB, & Moura LMVR (2022). A replicable, open-source, data integration method to support national practice-based research & quality improvement systems. Epilepsy Research, 186, 107013. 10.1016/j.eplepsyres.2022.107013 [PubMed: 35994859]

Fonferko-Shadrach B, Lacey AS, Roberts A, Akbari A, Thompson S, Ford DV, Lyons RA, Rees MI, & Pickrell WO (2019). 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, 9(4), e023232. 10.1136/bmjopen-2018-023232

Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, Qiu J, Yao Y, Zhang A, Zhang L, Han W, Huang M, Jin Q, Lan Y, Liu Y, Liu Z, Lu Z, Qiu X, Song R, … Zhu J (2021). Pre-trained models: Past, present and future. AI Open, 2, 225–250. 10.1016/j.aiopen.2021.08.002

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, & Stoyanov V (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach (arXiv:1907.11692). arXiv. 10.48550/arXiv.1907.11692

Munger Clary H, Josephson SA, Franklin G, Herman ST, Hopp JL, Hughes I, Meunier L, Moura LMVR, Parker-McFadden B, Pugh MJ, Schultz R, Spanaki MV, Bennett A, & Baca C (2022). Seizure Frequency Process and Outcome Quality Measures: Quality Improvement in Neurology. Neurology, 98(14), 583–590. 10.1212/WNL.0000000000200239 [PubMed: 35379694]

Patel AD, Baca C, Franklin G, Herman ST, Hughes I, Meunier L, Moura LMVR, Clary HM, Parker-McFadden B, Pugh MJ, Schultz RJ, Spanaki MV, Bennett A, & Josephson SA (2018). Quality improvement in neurology: Epilepsy Quality Measurement Set 2017 update. Neurology, 91(18), 829–836. 10.1212/WNL.0000000000006425 [PubMed: 30282773]

Pevy N, Christensen H, Walker T, & Reuber M (2021). Feasibility of using an automated analysis of formulation effort in patients' spoken seizure descriptions in the differential diagnosis of epileptic and nonepileptic seizures. Seizure - European Journal of Epilepsy, 91, 141–145. 10.1016/j.seizure.2021.06.009 [PubMed: 34157636]

Reiter E, & Dale R (1997). Building applied natural language generation systems. Natural Language Engineering, 3(1), 57–87. 10.1017/S1351324997001502

Saito T, & Rehmsmeier M (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. 10.1371/journal.pone.0118432 [PubMed: 25738806]

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, & Kattan MW (2010). Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology, 21(1), 128–138. 10.1097/EDE.0b013e3181c30fb2 [PubMed: 20010215]

Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M, & STROBE Initiative. (2014). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. International Journal of Surgery (London, England), 12(12), 1500–1524. 10.1016/j.ijsu.2014.07.014 [PubMed: 25046751]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, & Polosukhin I (2017). Attention Is All You Need (arXiv:1706.03762). arXiv. 10.48550/arXiv.1706.03762

Xie K, Gallagher RS, Conrad EC, Garrick CO, Baldassano SN, Bernabei JM, Galer PD, Ghosn NJ, Greenblatt AS, Jennings T, Kornspun A, Kulick-Soper CV, Panchal JM, Pattnaik AR, Scheid BH, Wei D, Weitzman M, Muthukrishnan R, Kim J, … Roth D (2022). Extracting seizure frequency from epilepsy clinic notes: A machine reading approach to natural language processing. Journal of the American Medical Informatics Association: JAMIA, 29(5), 873–881. 10.1093/jamia/ocac018 [PubMed: 35190834]

Xie K, Litt B, Roth D, & Ellis CA (2022). Quantifying Clinical Outcome Measures in Patients with Epilepsy Using the Electronic Health Record. Proceedings of the 21st Workshop on Biomedical Language Processing, 369–375. 10.18653/v1/2022.bionlp-1.36

Highlights

- Monitoring seizure control metrics is key for patients with epilepsy
- Manual abstraction of seizure control metrics from EHR notes is laborious
- Large Language Models can extract seizure control metrics from EHR notes
- Seizure control metrics extraction from EHR enables large-scale EHR research

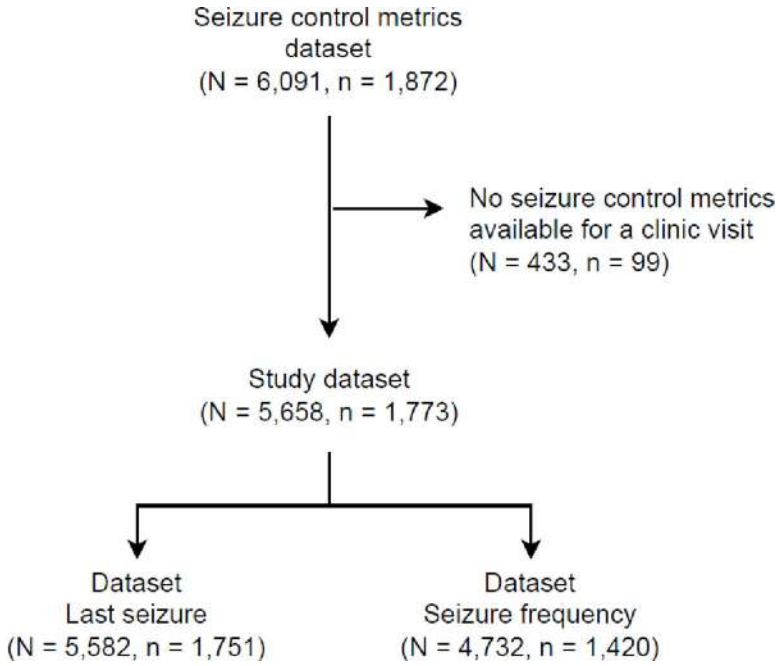

Figure 1.
Inclusion and exclusion criteria. N is the number of visits and n the number of patients.

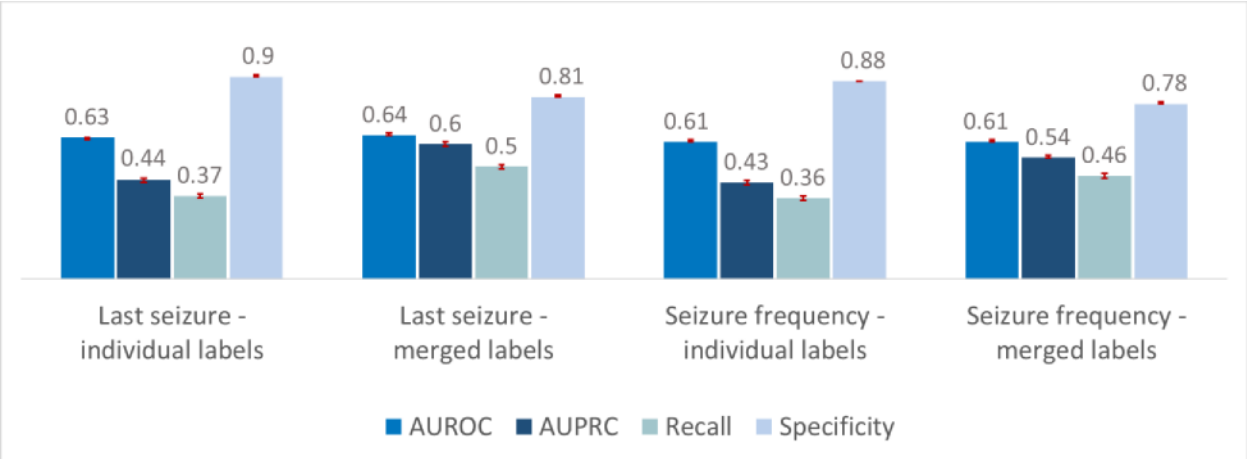

Figure 2.

Micro average modeling performance for date of last seizure and seizure frequency, with categorical labels, performed for 1000 bootstrapping iterations to calculate 95% confidence intervals.

AUROC – Area under the receiver operating curve; AUPRC – Area under the precision-recall curve.

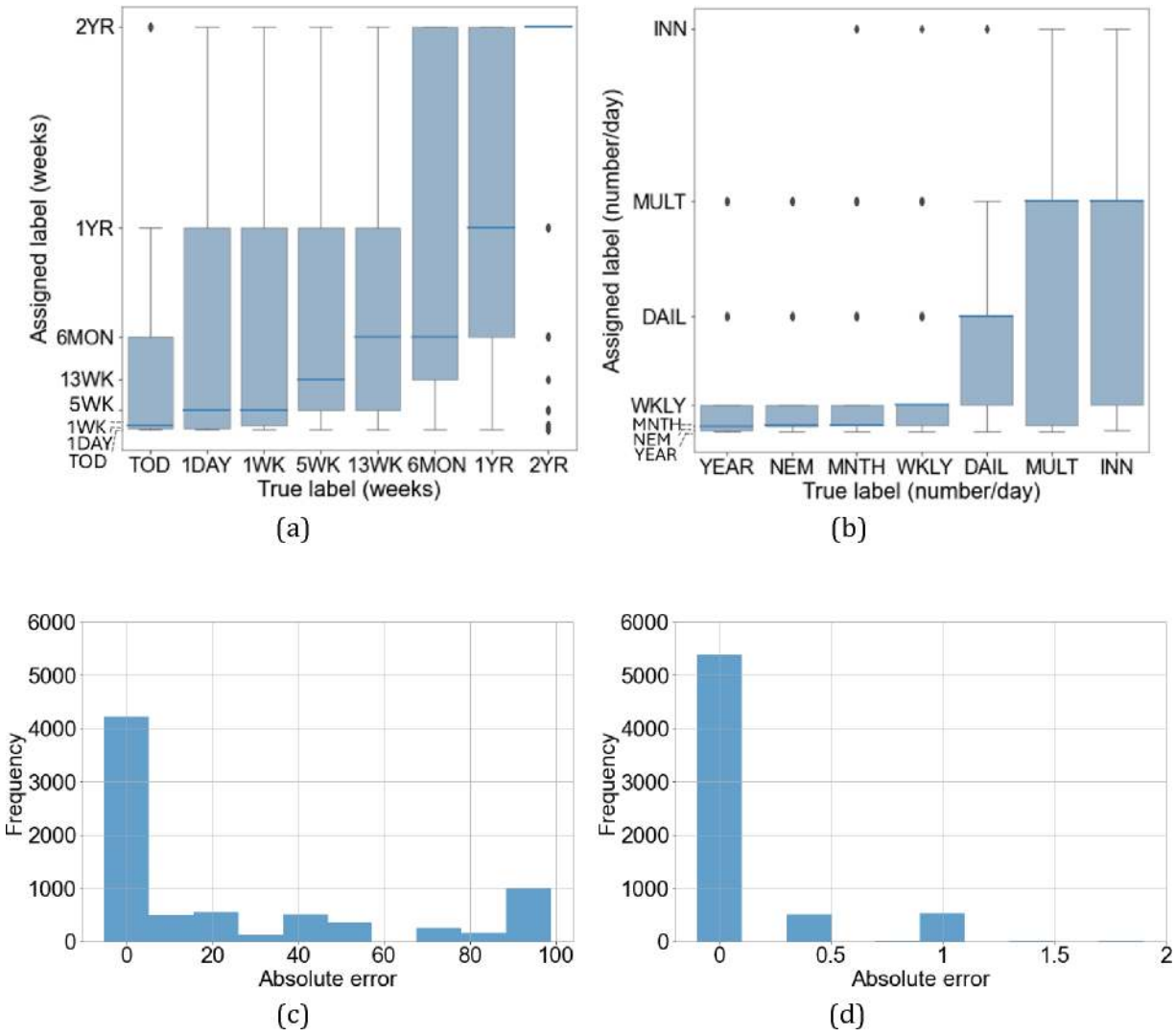

Figure 3.

Distribution of assigned vs true labels and the absolute error in the classification for (a)(c) date of last seizure and (b)(d) seizure frequency.

Table 1.

Prompts and corresponding text extraction method used for the prompt/question-answering task.

Extraction	Prompt	Method
Date of last seizure	*Last_seizure:* "When was the patient last seizure?" *Last_event:* "When was the patient last event?" *Recent_seizure:* "When was the patient most recent seizure?" *Recent_event:* "When was the patient most recent event?"	RoBERTa – *indirect approach*
	Last seizure: "Last seizure" *Last sz:* "Last sz" *Last convulsion:* "Last convulsion" *Last event:* "Last event"	Regexes – *semi-direct approach*
Seizure frequency	*Often_event:* "How often does the patient have seizures?" *Often_seizure:* "How often does the patient have seizures?" *Frequency_seizure:* "What is the seizure frequency?" *Frequency_event:* "What is the event frequency?"	RoBERTa – *indirect approach*
	Seizure frequency: "Seizure frequency" or "sz frequency"	Regexes – *semi-direct approach*

Table 2.

Specifications for label assignments in each field relevant to seizure control information.

Field	Label
Date of last seizure	TOD – Today 1DAY – 1 day to 6 days ago 1WK – 1 week to 4 weeks ago 5WK – More than 1 month to 3 months ago 13WK – More than 3 months to 6 months ago 6MON – More than 6 months to 12 months ago 1YR – More than 1 year to 2 years ago 2YR – More than 2 years ago
Seizure frequency since last visit	INN – Innumerable (i.e. ≥10 per day on most days) MULT – Multiple per day (i.e. 4 days per week with ≥2 seizures) DAIL – Daily (i.e. 4 or more days in the past week) WKLY – Weekly but not daily (i.e. 1 – 3 in the past week) MNTH – Monthly but not weekly (i.e. 1 – 3 in the past month) NEM – At least once per year, but not every month (i.e. 10 or fewer in past 12 months) YEAR – Less than once per year

Table 3.

Characteristics of the study cohort with reported seizure control metrics.

Characteristic	Study cohort (n = 1,773, N = 5,658)
Age[(a)] (years, mean (SD))	42.2 ± 16.4
Sex, n (%)	
Female	1,017 (57.4)
Race, n (%)	
Black or African American	109 (6.1)
Other [(b)]	229 (13.0)
White	1,435 (80.9)
Ethnicity, n (%)	
Hispanic	134 (7.6)
Unknown	131 (7.4)
Non-Hispanic	1,508 (85.0)
Epilepsy diagnosis, n (%)	1,773 (100.0)

The number of patients is represented by n and the number of visits by N.

[(a)] Age at baseline for the first visit in the study period.

[(b)] Other includes unknown, American Indian or Alaska Native and Asian.