Quantifying Clinical Outcome Measures in Patients with Epilepsy Using the Electronic Health Record

Kevin Xie[1,2], **Brian Litt**[1,2,3], **Dan Roth**[4] and **Colin A. Ellis**[2,3]
[1]Department of Bioengineering, University of Pennsylvania
[2]Center for Neuroengineering and Therapeutics, University of Pennsylvania
[3]Department of Neurology, Perelman School of Medicine, University of Pennsylvania
[4]Department of Computer and Information Sciences, University of Pennsylvania
kevinxie@seas.upenn.edu

Abstract

A wealth of important clinical information lies untouched in the Electronic Health Record, often in the form of unstructured textual documents. For patients with Epilepsy, such information includes outcome measures like Seizure Frequency and Dates of Last Seizure, key parameters that guide all therapy for these patients. Transformer models have been able to extract such outcome measures from unstructured clinical note text as sentences with human-like accuracy; however, these sentences are not yet usable in a quantitative analysis for large-scale studies. In this study, we developed a pipeline to quantify these outcome measures. We used text summarization models to convert unstructured sentences into specific formats, and then employed rules-based quantifiers to calculate seizure frequencies and dates of last seizure. We demonstrated that our pipeline of models does not excessively propagate errors and we analyzed its mistakes. We anticipate that our methods can be generalized outside of epilepsy to other disorders to drive large-scale clinical research.

1 Introduction

The Electronic Health Record (EHR) is a longitudinal catalog that describes patient visits, conditions, treatments, and well-being; thus, the EHR has significant potential for use in clinical informatics. Unfortunately, much of the data in the EHR is stored as unstructured text in the form of hand-typed doctor's notes, which makes rapid information extraction traditionally difficult. However, recent developments in neural models, namely Transformers (Vaswani et al., 2017) like BERT (Devlin et al., 2019), have opened up exciting new avenues of research.

Such developments have been applied to Epilepsy, a neurological disease characterized by recurrent unprovoked seizures. In epilepsy, seizure frequency and the date a seizure most recently occurred on are among the most important clinical outcome measures for patients. In Xie et al. (2022), we previously used specially finetuned Bio_ClinicalBERT (Alsentzer et al., 2019) and RoBERTa (Liu et al., 2019) models to extract, with human-like performance, sentences with a patient's seizure frequency and date of last seizure from clinical progress notes. These sentences contain temporal information and can thus be considered time expressions (timex).

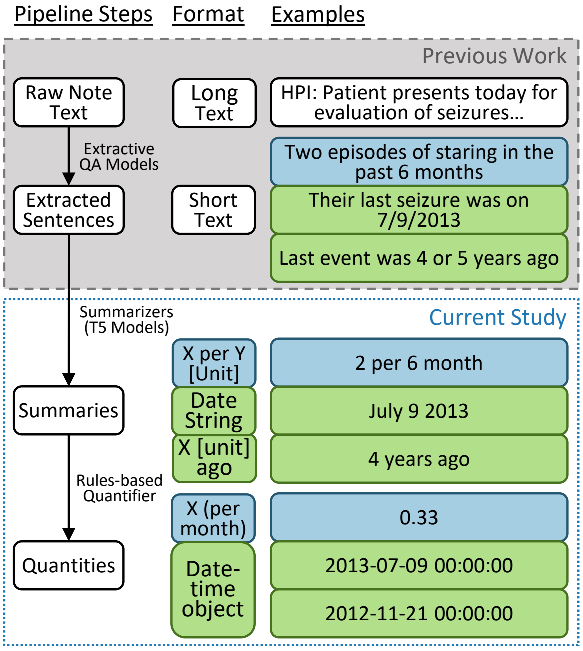

Figure 1: Schematic figure illustrating overall pipeline. Extractive question-answering models identify sentences containing seizure frequency and date of last seizure. These sentences are summarized into standardized formats. Quantities are extracted from these summaries using rules-based quantifiers. Items in blue are for seizure frequencies, while items in green are for dates of last seizure. Items in the grey background indicate work previously done in Xie et al. (2022).

However, such timex only simplify the problem of extracting such information from a document to extracting such information from a sentence, and

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 369–375
May 26, 2022. ©2022 Association for Computational Linguistics

are not yet usable in a quantitative manner. In this study, we developed a pipeline that extends our previous work and normalizes these timex (Figure 1[1]). We used neural text summarization models to convert the extracted information into a standardized format, and then applied a simple rules-based quantifier to calculate a quantitative seizure frequency (in seizures per month), or quantitative datetime object. Our approach required minimal annotation and preparation, and can be easily generalized to other similar tasks.

2 Methods

This retrospective study was approved by the Institutional Review Board of the University of Pennsylvania with a waiver of informed consent.

In Xie et al. (2022), we finetuned Bio_ClinicalBERT (Alsentzer et al., 2019) and RoBERTa (Liu et al., 2019) on a combination of public datasets and proprietary clinical notes to extract sentences with seizure frequency and dates of last seizure from clinical notes. We framed this task as an extractive question-answering problem, where we asked the model to identify statements that answered the questions "How often does the patient have seizures," and "When was the patient's most recent seizure?" We demonstrated that our models achieved human-like performance relative to clinicians and researchers working in epilepsy-related research.

Quantifying seizure frequency and dates of last seizures from these sentences is therefore a timex normalization task, which seeks to convert a timex statement like "Their last seizure was on 7/9/2013" into the datetime object 2013-07-09 00:00:00. More difficultly, seizure frequency and date of last seizure are represented in a number of non-standardized ways by clinicians, precluding the use of simple rules-based quantification. We characterize broad categories and provide illustrative examples of such representations in Table 1. Note that such representations are often encapsulated by surrounding text (e.g. "They continue to have nocturnal convulsive seizures twice per week"), and that each category has internal variance (e.g. "seizure weekly" vs. "seizures once per week"). To accommodate these representations, we split our timex normalization process into two steps: simplification and quantification.

Frequency Format	Example
Classical	"weekly basis", "twice per week"
Implied	"first day of their menses"
Calendar	"January: 1, February: 3, ..."
Timepoint	"Since last visit... 3 seizures"

Last Seizure Format	Example
Explicit	"Last seizure was 3/2012"
Implicit	"Seizure free since 2001"
Timepoint	"...2 or 3 years ago"

Table 1: Broad categories of seizure frequency and date of last seizure formats with corresponding examples.

We first attempted to simplify each sentence into a standardized format: "X per Y [day/month/year/visit]" (e.g. "1 per 1 week") for seizure frequencies, and "[Month] [Day] [Year]" or "X [day/month/year] ago" (e.g. "January 2012" or "3 years ago") for date of last seizure. We frame this task as an abstractive text summary problem: given a sentence containing a seizure frequency or date of last seizure, we summarize the main component of the sentence, the frequency or date, into a standardized template. We manually annotated the 1,000 sentences of seizure frequency and 1,000 sentences of the date of last seizure previously generated by our models in Xie et al. (2022) with the formatted summaries; for example, "Two episodes of staring in the past 6 months" was annotated with "2 per 6 months", and "Their last seizure was on 7/9/2013" was annotated with "July 9 2013". We then split them into training and testing sets, with 700 sentences for training, and 300 for testing. We also created concrete values for subjective statements (i.e. "many", "few", etc...) (Appendix A).

We finetuned two T5-large models (Raffel et al., 2020) using Huggingface (Wolf et al., 2020), on the training sets and made predictions on the test sets. One T5-large model summarized sentences of seizure frequency, while the other summarized sentences of last seizure. We used Huggingface's default parameters for text summarization and did not perform any hyperparameter optimization.

We then developed a rules-based quantifier that normalizes a frequency summary into a numerical value, and converts a date summary into a datetime object. For summaries of seizure frequency, we take the "X" value in "X per Y

[1] Our examples are date-shifted and gender neutralized when applicable to preserve patient privacy and HIPAA compliance

Sentence	**Summary**	**Quantity**
Seizures persisted throughout their life, approximately once a year	1 per 1 year	0.0833
... Jan 5 clusters, Feb 10 clusters, March 4 clusters, April 8 clusters	4 per 6 month*	6.75
Two episodes of staring in the past 6 months	2 per 6 months	0.333
Their last seizure was on 7/9/2013	July 9 2013	2013-07-09 00:00:00
Last event was 4 or 5 years ago	4 years ago	2012-11-21 00:00:00
Not had any seizures since 2005	2005	2005-01-01 00:00:00

Table 2: Examples of the summary and quantification processes to quantify sentences of seizure frequency and date of last seizure.
*Note: the seizure calendar sentence's summary was incorrect, but the final quantity was corrected using the rules-based quantifier for seizure calendars.

[day/month/year/visit]" as the numerator, and convert the "Y" value using the time unit given in "[day/month/year/visit]" into a suitable denominator to have units of "seizures per month." If the timeframe involved the previous visit ("per Y visit"), we would attempt to search for a record of the patient's last visit in our dataset and calculate the number of months that have passed; if no such record could be found, the quantifier would insert a placeholder statement for future analysis when such information would be available. For summaries of date of last seizure, we first determine if the summary was of the "ago" form, in which case we subtract the specified number of day, months, or years from the date the note was written. Otherwise, we apply a series of logical steps to quantify the summary into a Python datetime object. If only a month and day were given, we assume that the year was either the same year that the note was written, or the previous year, depending on if the resultant date using the same year was in the future of the date the note was written. In both quantifiers, we assume that there are 365 days or 12 months in a year, 7 days in a week, and 30.4167 days or 4.3452 weeks in a month.

We also created a rules-based quantifier specifically for the seizure calendars, as the summarizer was unable to produce an accurate summary of this format of frequency. This seizure calendar quantifier identifies a sentence as a seizure calendar if it has at least two months, and at least two numbers. It then associates a month to its number of seizures by assuming that the number of seizures either directly follows the month in the text (e.g. "January: 1"), or precedes the month within three words (e.g. "1 seizure in January")." It counts the number of months and accumulates the number of seizures in that time span to calculate a monthly seizure frequency. Table 2 provides some examples of the overall process.

We manually calculated the accuracy of each step of our approach in an all-or-nothing approach by comparing a statement to its downstream summary or quantity; a step was correct only if both its format and value given the context were correct.

3 Results

We finetuned our T5 models for text summarization using a training set of 700 annotations, and a testing set of 300 annotations. To determine how much error we were propagating through our pipeline, we calculated the accuracy of each step in our method using the testing set (Table 3). We counted the number of accurate sentences from medical notes (performed previously in Xie et al. (2022)), summaries (accounting for both correct value and format) from sentences, and quantities from summaries. Note that for this calculation, we considered each step of the process as independent from the others; for example, a summary could be correct given a sentence, even if that sentence itself was incorrect relative to the original note text. We also determined the overall accuracy as the number of examples where all of these steps were correct. With at least 96% accuracy, it is evident that our summarizers produced consistent representations of seizure frequency and date of last seizure in the desired format. Meanwhile, our perfect quantification accuracy validates our use of text summaries as an intermediate step - because all seizure frequencies and dates of last seizure have been consistently converted into their own respective formats, it is

	Sentence Accuracy	Summary Accuracy	Quantity Accuracy	Overall Accuracy
Seizure Frequency	0.893	0.963	1.000	0.880
Date of Last Seizure	0.863	0.987	1.000	0.857

Table 3: Accuracies of the extracted sentences containing seizure frequencies or dates of last seizure from raw clinical note text (described previously in Xie et al. (2022)), the summary of such sentences in the standardized format, and the quantification of the summaries into quantities. The overall accuracy denotes how often every step of this process was correct. For calendar-type seizure frequencies, overall accuracy ignores the summary step, as this was always incorrect, and instead takes into account the seizure calendar quantifier.

Reason	Times Erred (Seizure Frequency)	Times Erred (Last Seizure)
Competitive Temporal Statements "Since last visit: ... one ... seizure in the past year" "On the same day as their last appointment ..."	**2**	**2**
No Temporal Reference "They think they only had two ... seizures" "Two weeks later they had another seizure"	**2**	**2**
Using Month as Value "Since 4/2012 they have had a few seizures" "Since last office visit, they have had seizure 8/12/16"	**2**	**0**

Table 4: Types of errors that occurred during the summary process.

highly unlikely that some unforeseen representation will be able to break the quantifiers' rules.

Finally, we attempted to identify patterns of errors in our incorrect summaries. We manually catalogued these errors for both sentences of seizure frequency and dates of last seizure, and determined potential reasons for such problems (Table 4). The first category was for sentences with competitive time modalities, e.g. "Since last visit: they report one possible seizure in the past year". Here the summary could either use "since last visit" or "past year" as its temporal unit for a seizure frequency; in this particular example, the model chose to use "since last visit", when "past year" would have been more appropriate. Similarly, there were cases when a temporal reference point was not available, such as this sentence of a date of last seizure: "Two weeks later they had another seizure." In this case, it is unknown when exactly "two weeks later" is referring to. This is reflected in the model's summary for this example - "2 weeks later". Though in some sense correct, this summary did not follow the desired format, namely because there was not enough information, even for a human, to fit it within the specified style[2]. Finally, some cases where seizure frequencies with dates were written out in numerical format resulted in the model pulling elements of those dates out as part of the frequency itself. For example "Since last office visit, they have had seizure 8/12/16" was summarized as "8 per 1 visit", but "8/12/16" instead refers to the date at which their seizure occurred; the correct summary should have been "1 per 1 visit".

4 Discussion

In this study, we normalized timex containing seizure frequency and date of last seizure by simplifying them with text summarization models, and applying simple rules-based quantifiers to extract quantitative outcome measures for patients with epilepsy. We demonstrated that this pipeline can accurately calculate quantitative seizure frequencies and dates of last seizure. Though applied specifically to epilepsy, our methods are not constrained just to neurological disorders, and can be easily adapted to other medical conditions as well. Our findings pave the way for large-scale clinical informatics research through extracting and quantifying textual information from the EHR.

Our full pipeline, including our previous work from Xie et al. (2022), extracts timex from clinical documents, simplifies them using neural models,

[2] The quantifier correctly flagged this summary as anomalous and did not produce a quantity.

and normalizes them with rules-based methods to obtain quantitative outcome measures. The overall process is reminiscent of other temporal understanding studies. For example, Ning et al. (2018) developed a pipeline for temporal understanding that involves a Begin-Inside-Outside (BIO) tagging scheme with machine learning to extract timex, and a rules-based method to normalize them. Meanwhile, Ding et al. (2021) formulated timex normalization as a sequence of operations that selects and applies normalization rules, and Miller et al. (2015) extracted timex from clinical text using machine learning-based BIO taggers on two clinical datasets.

Additionally, to our knowledge, we are the first to use neural text summarization as an intermediate step to simplify variable timex into a standardized template for easy rules-based quantity extraction in the clinical domain. However, similar approaches exist in other domains and tasks. For example, Lourentzou et al. (2019) used a seq-to-seq model to normalize the often complex and non-standard text found in social media into more standard forms. Additionally, Vale et al. (2018) tested how various sentence simplification methods improved the informativeness of extractive text summarization methods, while Che et al. (2015) compressed sentences in a manner that simplified the sentence but preserved its sentiment as a preprocessing step for aspect-based sentiment analysis.

Our categories of errors are also in line with what has been seen in the literature for Transformers. For example, Sulem et al. (2021) found that in extractive question-answering tasks, BERT models showed remarkably lower performance on competitive I-Don't-Know questions (where a plausible but incorrect answer of the correct type exists in the context), mirroring our summarization errors when competitive time frames were presented.

Our study does have limitations. First and foremost, our methodology was developed using data from a single institutional healthcare center. While we used a neural summarizer in the hopes of improving overall generalizability to the various ways of representing outcome measures in text, it is still possible that the summarizer will fail to generalize to text from other health centers. We are actively evaluating of our methods at a collaborating institution to access this effect. Additionally, 21 of 22 summaries that involved previous visits could not be actively quantified with this dataset, as the date of the previous visit did not exist in the 300 test notes. This can easily be corrected by performing a larger longitudinal study across our patients that would allow us to track them through their visits.

5 Conclusions

We created a generalized two-step system that rapidly and accurately extracts and quantifies seizure frequency and date of last seizures. We used the T5 model to create standardized summaries of sentences of these outcome measures, and then applied a rules-based algorithm to extract and quantify the desired information. We anticipate that our methods can be used to quantify important clinical outcome measures not only for patients with epilepsy, but other disorders as well, allowing for large-scale clinical research in the future.

Acknowledgements

This research was funded by the National Institute of Neurological Disorders and Stroke 1DP1 OD029758; by the Mirowski Family Foundation; and by contributions from Jonathan and Bonnie Rothberg. CAE is supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health Award Number K23NS121520; by the American Academy of Neurology Susan S. Spencer Clinical Research Training Scholarship; and by the Mirowski Family Foundation. DR's work was partially funded by the Office of Naval Research Contract N00014-19-1-2620.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu. 2015. Sentence compression for aspect-based sentiment analysis. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(12):2111–2124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wentao Ding, Jianhao Chen, Jinmao Li, and Yuzhong Qu. 2021. Automatic rule generation for time expression normalization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3135–3144, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ismini Lourentzou, Kabir Manghnani, and ChengXiang Zhai. 2019. Adapting sequence to sequence models for text normalization in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):335–345.

Timothy Miller, Steven Bethard, Dmitriy Dligach, Chen Lin, and Guergana Savova. 2015. Extracting time expressions from clinical text. In *Proceedings of BioNLP 15*, pages 81–91, Beijing, China. Association for Computational Linguistics.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Elior Sulem, Jamaal Hay, and Dan Roth. 2021. Do we know what we don't know? studying unanswerable questions beyond SQuAD 2.0. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4543–4548, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rafaella Vale, Rafael Lins, and Rafael Ferreira. 2018. Assessing sentence simplification methods applied to text summarization. In *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 49–54.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kevin Xie, Ryan S Gallagher, Erin C Conrad, Chadric O Garrick, Steven N Baldassano, John M Bernabei, Peter D Galer, Nina J Ghosn, Adam S Greenblatt, Tara Jennings, Alana Kornspun, Catherine V Kulick-Soper, Jal M Panchal, Akash R Pattnaik, Brittany H Scheid, Danmeng Wei, Micah Weitzman, Ramya Muthukrishnan, Joongwon Kim, Brian Litt, Colin A Ellis, and Dan Roth. 2022. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. *Journal of the American Medical Informatics Association*. Ocac018.

A Subjective Statement Values

Statement	Value
"Couple"	2
"Few"	3
"Several"	4
"Multiple"	4
"Many"	5

Table 5: Values for subjective statements. Values were chosen by consensus of the authors.