HHS Public Access

Author manuscript

Epilepsy Res. Author manuscript; available in PMC 2023 January 03.

Published in final edited form as:

Epilepsy Res. 2022 October ; 186: 107013. doi:10.1016/j.eplepsyres.2022.107013.

A replicable, open-source, data integration method to support national practice-based research & quality improvement systems

Marta Fernandes[a,b,c,*], **Maria A. Donahue**[a,b,d], **Dan Hoch**[a,b], **Sydney Cash**[a,b], **Sahar Zafar**[a,b], **Claire Jacobs**[a,b], **Mackenzie Hosford**[a,b], **P. Emanuela Voinescu**[b,e], **Brandy Fureman**[f], **Jeffrey Buchhalter**[g], **Christopher Michael McGraw**[a,b,1], **M. Brandon Westover**[a,b,c,h,1], **Lidia M.V.R. Moura**[a,b,d,1]

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

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

[c]Clinical Data Animation Center (CDAC), MGH, Boston, MA, United States

[d]The NeuroValue Lab, MGH, Boston, MA, United States

[e]Department of Neurology, Division of Epilepsy, Division of Women's Health, Brigham and Women's Hospital, Boston, MA, United States

[f]Epilepsy Foundation of America, United States

[g]Department of Pediatrics, University of Calgary School of Medicine, Calgary, Canada

[h]McCance Center for Brain Health, MGH, Boston, MA, United States

Abstract

Objectives: The Epilepsy Learning Healthcare System (ELHS) was created in 2018 to address measurable improvements in outcomes for people with epilepsy. However, fragmentation of data systems has been a major barrier for reporting and participation. In this study, we aimed to test the feasibility of an open-source Data Integration (DI) method that connects real-life clinical data to national research and quality improvement (QI) systems.

Methods: The ELHS case report forms were programmed as EPIC SmartPhrases at Mass General Brigham (MGB) in December 2018 and subsequently as EPIC SmartForms in June 2021 to collect actionable, standardized, structured epilepsy data in the electronic health record (EHR) for subsequent pull into the external national registry of the ELHS. Following the QI methodology in the Chronic Care Model, 39 providers, epileptologists and neurologists, incorporated the ELHS SmartPhrase into their clinical workflow, focusing on collecting diagnosis of epilepsy, seizure type

[*]Corresponding author at: Department of Neurology, Massachusetts General Hospital (MGH), Boston, MA, United States. mbentofernandes@mgh.harvard.edu (M. Fernandes).

[1]Co-senior authors

Declarations of Interest

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.eplepsyres.2022.107013.

according to the International League Against Epilepsy, seizure frequency, date of last seizure, medication adherence and side effects. The collected data was stored in the Enterprise Data Warehouse (EDW) without integration with external systems. We developed and validated a DI method that extracted the data from EDW using structured query language and later preprocessed using text mining. We used the ELHS data dictionary to match fields in the preprocessed notes to obtain the final structured dataset with seizure control information. For illustration, we described the data curated from the care period of 12/2018–12/2021.

Results: The cohort comprised a total of 1806 patients with a mean age of 43 years old (SD: 17.0), where 57% were female, 80% were white, and 84% were non-Hispanic/Latino. Using our DI method, we automated the data mining, preprocessing, and exporting of the structured dataset into a local database, to be weekly accessible to clinicians and quality improvers. During the period of SmartPhrase implementation, there were 5168 clinic visits logged by providers documenting each patient's seizure type and frequency. During this period, providers documented 59% patients having focal seizures, 35% having generalized seizures and 6% patients having another type. Of the cohort, 45% patients had private insurance. The resulting structured dataset was bulk uploaded via web interface into the external national registry of the ELHS.

Conclusions: Structured data can be feasibly extracted from text notes of epilepsy patients for weekly reporting to a national learning healthcare system.

Keywords

Epilepsy; International League Against Epilepsy; Seizure control; Quality Improvement; Learning Health System; Text analytics

1. Introduction

Digital technology plays an increasingly critical role in healthcare organizations, enabling the reduction in the reporting burden for health care staff and facilitating the effective use of data for practice improvement, clinical decision making, and population management (Bell et al., 2018; Meskó et al., 2017; Thimbleby, 2013). Especially in health information systems with the transition to electronic health records (EHR), the development of digital tools to help collect and report health data has proved to be a key need (World Health Organization, 2019b). However, there are many barriers to data driven healthcare improvement and several approaches have emerged including Learning Healthcare Systems (LHS) (Enticott et al., 2021). LHS use continuous Quality Improvement (QI) strategies to improve care over time (Morain et al., 2017). LHS models embed healthcare data-driven research, integrating multidisciplinary expertise to deliver improved health (Budrionis and Bellika, 2016; McLachlan et al., 2018; Menear et al., 2019; Scobie and Castle-Clarke, 2019), through data collection and analysis at the point of care to inform clinical decision making (Scobie and Castle-Clarke, 2019; Teede et al., 2019), thereby improving healthcare and outcomes (Enticott et al., 2021). Clinicians use common data elements (CDEs), which consist of standardized questions and answer choices that allow aggregation, analysis, and comparison of observations from multiple sources (Loring et al., 2011), and are foundational for LHS.

Even though LHS environments are producing successful impact across multiple continents and settings (Enticott et al., 2021), very few LHSs focus on neurological disorders (Bindman et al., 2018). The Epilepsy Learning Healthcare System (ELHS) was created in 2018 to improve epilepsy care at a system level by addressing measurable improvements in outcomes for people with epilepsy (Donahue et al., 2021). Epilepsy is a chronic disease accounting for 0.5% of the global burden of disease, a time-based measure that combines years of life lost due to premature mortality and time lived in less than full health (World Health Organization, 2019a). Epilepsy affects 50–70 million people of all ages worldwide (Ngugi et al., 2010), being the fourth most common neurological disorder globally (Epilepsy Foundation, 2020b; Scheffer et al., 2017). Epilepsy has significant economic implications in terms of health-care needs, premature death and lost work productivity (Ngugi et al., 2010; World Health Organization, 2019a). Despite significant pharmacological and technological advances in the field of epilepsy, at least 3 out of 10 people with epilepsy continue to have seizures because available treatments do not completely control their seizures (Epilepsy Foundation, 2020a; Forsgren et al., 2005).

ELHS goals include the improvement of care processes and clinical outcomes by utilizing QI methodology in the context of a Chronic Care Model (Margolis et al., 2013). Once ELHS standardized epilepsy care metrics are being collected reliably for a vast number of patients, in addition to directly improving outcomes via QI methodology, the large numbers will facilitate comparative effectiveness studies and identification of patients for large real-world clinical trials. Seizure frequency documentation and screening for barriers to medication adherence were selected by ELHS as the first two network-wide interventions (Donahue et al., 2021).

The ELHS case report forms (CRFs) were designed to collect information in a standardized manner that is essential to patient care, useful for QI and feasible within the clinic workflow. The ELHS CRFs were programmed as a SmartPhrase in the EPIC EHR software at Mass General Brigham (MGB) in December 2018 to collect actionable, standardized, structured epilepsy data in the EHR for data integration (DI) and subsequent pull into the national registry of the ELHS. Epileptologists and neurologists incorporated the ELHS SmartPhrase into their clinical workflow, focusing in collecting seizure control information, which was stored as semi-structured text data in the MGB system database. Although the SmartPhrase allows physicians to enter data in a structured manner, the backend database stores the data as free text. Thus, to make use of this data for research and quality improvement processes, it is necessary to create an automated extraction method (Elbattah et al., 2021).

Text Analytics, or Text Mining, is generally defined as “the methodology followed to derive quality and actionable insights from textual data” (Sarkar, 2019). Text Mining represents a field of techniques and technologies that include machine learning, natural language processing (NLP), and information retrieval. Leveraging the power of text mining permits transformation of semi-structured and unstructured text data into structured information from which powerful insights can be derived. Thus, in this paper we present a text mining methodology to extract seizure control information from the ELHS EHR SmartPhrase and analysis of this data. Using our DI method, we automated the data mining, preprocessing, and exporting of the structured dataset into a local database, to be accessible on a weekly

basis to clinicians and quality improvers. We uploaded this information to the national registry of ELHS, thereby promoting its development and growth.

2. Methods

2.1. Study overview

The ELHS developed CRFs to collect standardized epilepsy care related data (Donahue et al., 2021). The ELHS CRFs were programmed as EPIC SmartPhrases at MGB in December 2018 and subsequently as EPIC SmartForms in June 2021 to collect actionable, standardized, structured epilepsy data in the EHR for subsequent pull into the external national registry of the ELHS. Following the QI methodology in the Chronic Care Model (Margolis et al., 2013), a total of 39 epileptologists and neurologists incorporated the ELHS SmartPhrase into their clinical workflow, focusing on collecting seizure type according to the International League Against Epilepsy (ILAE) seizure classification (Scheffer et al., 2017), seizure frequency, date of last seizure and anti-seizure medication (ASM) adherence and side effects. ASM information fields were only added in a later version of the SmartPhrase, introduced in July 2020.

The collected data was stored in the MGB Enterprise Data Warehouse (EDW). The challenge consisted of the extraction of ELHS SmartPhrases data stored in EDW as text and the transformation of this data into a structured dataset so that it could be uploaded to the external national registry of the ELHS. The real-life process and challenges with clinical data flow associated with patients visits at the epilepsy clinics are illustrated in Fig. 1. By automating the process of clinical notes extraction and preprocessing – as illustrated in Fig. 2 – data analysis such as causal inferences and weekly QI dashboards can be potential applications.

Data was extracted from the hospital electronic medical record under a research protocol approved by the MGB Institutional Review Board; a waiver of informed consent was obtained. Clinical data were retrospectively analyzed for a cohort of adult patients (≥ 18 years old). The cohort included 1806 patients seen at the Massachusetts General Hospital (MGH, n = 1371) and Brigham and Women's Hospital (BWH, n = 435) epilepsy clinics, between December 28th, 2018, to December 20th, 2021. Both hospitals use the MGB EPIC EHR software system. EHR data comprised clinical notes consisting of structured SmartPhrases (Supplementary Figure 1).

2.2. Data Collection and Processing

Data was acquired from MGB EDW using Microsoft Structured Query Language (SQL) Server Management Studio (SSMS). The methodology for notes preprocessing in order to obtain a structured dataset is presented in Fig. 2. All notes from the patients seen at MGH and BWH epilepsy clinics were extracted. Patient identification (PatientID and medical record number (MRN)), healthcare provider, department identification, date of visit, note identification (NoteID), note line number (LineNBR) and the text notes were queried from the EDW databases and exported to an excel file. An example of a text note extracted from EDW for preprocessing is presented in Supplementary Figure 2. The EHR clinical

notes were then processed using Python version 3.7. Notes were filtered to include only those pertaining to the structured ELHS SmartForms. These notes were sorted by NoteID and LineNBR, and then grouped by PatientID, MRN, DepartmentID, provider and visit date. Notes were then subjected to lowercasing, blank spaces were removed, and notes with less than one token (one word), were excluded. Once the notes were preprocessed, regular expressions (regexes) were applied to extract information regarding seizure control, namely if the patient has epilepsy, if the patient missed taking the ASM, if there are side effects from taking the ASM, or any barriers to adherence to the ASM, the seizure type according to the ILAE classification, the correspondent diagnostic certainty, date of last seizure since current visit and current seizure frequency. For each of these fields, a column was created with the correspondent categories assigned. Notes without categories assigned for at least one of the seizure control information fields, were removed.

The providers also documented one to four seizure types – a, b, c and d –, since a patient may have more than one type. Thus, for each seizure type, the provider selects the respective type, frequency, diagnostic certainty and date of last seizure. Usually, when a patient only has one type of seizure, the provider will only select the fields correspondent to seizure type “a”, while if the patient has two types of seizure, both seizure types “a” and “b” fields will be selected. And so on, until four types of seizure (a, b, c and d) are selected in case the patient has four seizure types.

We present examples of the regexes created, including the number of tokens captured for information extraction for each field in Supplementary Table 1. The number of tokens was set for each field based on the length of text of the corresponding categories. For example, for the field “Does the patient have epilepsy?” we captured the corresponding answer yes/no/unsure, with a total maximum length of 100 tokens (words). For special cases, including menus with multiple choices, the number of tokens was set to a maximum of 3000 since menus include multiple options. As an example, for the field “seizure type”, we have “Seizure type A ILAE CLASSIFICATION: Generalized [x] Tonic-clonic [] Absence typical [] Absence atypical [] Myoclonic absence [] Eyelid myoclonia [] Myoclonic [] Myoclonic atonic [] Myoclonic tonic [] Clonic [] Tonic Focal [] Focal aware without impairment of consciousness [] With observable motor components [] Involving subjective sensory or psychic phenomena only [] Focal with impairment of consciousness [] Focal evolving to bilateral convulsive seizure Unclassified [] Epileptic spasms (focal or generalized) [] Unclassified seizure type”. For this example, we captured all the seizure types displayed in the menu and the regex assigned the seizure type generalized tonic-clonic “GENTC”.

The regular expressions were designed in a flexible way. In case that a provider has indicated an actual calendar date for the field date of last seizure, instead of selecting one of the coded options (eg. 1WK – one week ago or 6MON – six months ago), the regex captured the date and assigned the correspondent category. This approach was also applied for the field “seizure frequency”. We provide an example with the text “Seizure type A current FREQUENCY since last visit with same provider: GTC - less than 2 years ago. Dejavu - unclear.”, in Supplementary Figure 3 (case 1), where the regex assigned the corresponding category “YEAR” (less than once per year). We also analyzed cases where there was a complete note for the patient visit at the epilepsy clinic, however no information could be

extracted, because the seizure control information fields consisted of " *** ", as presented in Supplementary Figure 3 (case 2). For these few special cases, the SmartPhrase was not correctly saved by the system at the time of clinic visit. The regexes were refined iteratively by assessing notes with special cases and for each field where no category was assigned.

2.3. Performance evaluation

The authors (MF, SZ, CMM, MBW, LMVRM) manually reviewed 100 randomly selected notes (20 cases per author), for a total of 2000 fields (100 notes × 20 fields for seizure control metrics). The seizure control metrics included: whether the patient has epilepsy, whether the patient missed taking the ASM, if there are side effects from taking ASM(s), if there are barriers to adherence to the ASM, and for the four seizure types (a/b/c/d) the type, the corresponding diagnostic certainty, date of last seizure, and seizure frequency. We compared the performance of the regexes-based approach with the manual review results. Performance measures included precision, recall and F1 score to measure the regex-based approach compared to the gold standard of manual review. Precision is defined as the proportion of the seizure control metric items extracted by the regex-based approach which are true, recall is the proportion of true seizure control metrics items extracted by the algorithm and F1 score is the unweighted harmonic mean of precision and recall: (2 × precision × recall)/ (precision + recall).

The cohort population further analyzed included adult patients residing in the US seen at the epilepsy clinics, with information regarding their seizure type and corresponding frequency, present in the seizure control structured dataset.

3. Results

3.1. Summary of patient population

The EHR clinical notes were extracted from EDW and preprocessed. The categories for each seizure control information field were assigned according to the ELHS specifications, using regexes to obtain a structured dataset. There were 2125 patients with seizure control information in the dataset. After excluding patients younger than 18 years old, non-US residents and those who lacked both seizure type and frequency, we were left with 1806 patients (Fig. 3).

The cohort of patients had a mean age of 43 years old (SD: 17), where 57% were female, 80% were white, and 84 % were non-Hispanic/Latino (Table 1). There were 5168 clinic visits logged by providers documenting both seizure type and frequency. During this period, providers documented 59 % patients having focal seizures, 35 % having generalized seizures and 6 % patients having other type. Of the cohort, 45 % patients had private insurance.

3.2. Performance results

From the manual review of 100 notes randomly selected, the experts identified 890 items in a total of 2000 fields (100 notes × 20 fields for seizure control metrics). The majority of notes were from patients with one type of epilepsy (seizure type a), with correspondent diagnostic certainty (diagnostic certainty a), date of last seizure (date of last seizure a), and

seizure frequency (seizure frequency a). Thus, the majority of fields for the second, third and fourth types of epilepsy, b, c and d, respectively, were left blank, since these were not identified in the notes.

The regex-based approach captured 887 items from the 890 items identified by the experts, as presented in Table 2. The category “Not Applicable” was identified by experts for the metrics of seizure type, seizure frequency and date of last seizure. This category was not included in the ELHS specifications for these metrics, as depicted in Supplementary Table 2, thus it was not coded in the algorithm. Among 100 notes randomly selected, there were 3 false negative cases. The overall performance was approximately 100 % (precision: 1, recall: 0.996, F-score: 0.998). The performance obtained is mainly explained by the fact that notes are mostly structured, and might not reflect the overall performance, since it considers only 2 % (100/5168) of the notes in the study.

3.3. Analysis of documented seizure control metrics

The number of patients with documented fields in the notes can be depicted in Fig. 4. Patients may have follow-up visits, thus the volume of notes is greater than the number of patients. For each patient, a category for a specific field may repeat (e.g. “YES” for the question “Does the patient have epilepsy?”) and for another field, such as seizure frequency, the category may vary for example, from “NEM” (at least once per year, but not every month) to “YEAR” (less than once per year).

From a total of 1806 patients with documented seizure type and frequency, almost all of the patients (N = 1804) had a category assigned for the field “Diagnostic certainty”, as well as for the field “Does the patient have epilepsy?” (N = 1801). The “Date of last seizure” field was also documented for the majority of patients (N = 1792). The fields less documented were the ones related to ASM, namely the fields: “In the past four weeks, how good of a job did the patient do at taking anti-seizure medicine?” (N = 1021), “Is the patient having any problems that are side effects of the medicine?” (N = 1018) and “Is the patient experiencing any barriers to adherence?” (N = 847). This is due to the fact that documentation of ASM was added in a later version of the SmartPhrase, and the first register date is July 24th, 2020. In a total of 1394 patients with notes prior to ASM documentation, approximately 73 % (N = 1021) have documentation of missed ASM, as observed in Fig. 4 (a).

We assessed the percentage of patients with reported seizure control metrics documented for the period before and after the implementation of ASM documentation, presented in Fig. 5. For the most reported fields, the percentage of patients with reported seizure control metrics was consistently ~ 100 %. The fields with medication related information documented was approximately 73 % for both missed medication (N = 1021) and side effects (N = 1018) and 61 % for barriers to adherence (N = 847).

Finally, we assessed the distribution of categories assigned by providers per field in the study cohort (Supplementary Table 3). The majority of patients had epilepsy (N = 1554, 86 %) and was assigned Generalized tonic-clonic (N = 481), Focal with impairment of consciousness (N = 147) and Focal evolving to bilateral convulsive seizure (N = 33) for the three seizure types a, b and c, respectively. Since the number of patients with a fourth seizure

type was reduced, we omitted the numbers for this seizure type. The majority of patients had a definite diagnostic – summary of evidence suggesting 100 % confidence level – for all three seizure types a (N = 1053), b (N = 344) and c (N = 65). For the majority, the date of last seizure types a, b and c (N = 504, N = 200 and N = 42) was more than 2 years prior to the visit; and the frequency since last visit for seizure types a, b and c (N = 561, N = 179 and N = 34) was less than once per year. For the fields related to ASM, when documented, the majority of patients did not miss more than two doses in a row of ASM (N = 590), did not present side effects to the ASM (N = 517) and did not experience any barriers to adherence (N = 492).

4. Discussion

4.1. Main findings

We collected standardized epilepsy care data in the EHR of patients seen at MGB. This data consisted of diagnosis of epilepsy, seizure type according to the ILAE classification, seizure frequency, date of last seizure and anti-seizure medication adherence and side effects. The data was extracted from EDW using structured query language and preprocessed using regexes. Using the ELHS data dictionary, we matched the fields in the preprocessed notes to organize the data into a structured format that we analyzed. We assessed the performance of our regex-based approach compared to experts' manual review of 100 random notes. The overall performance of the regex-based approach was approximately 100 % across metrics, which we concluded was due to the fact that notes are mostly structured. We found that the majority of patients with documented seizure type and frequency had a definite diagnostic certainty of epilepsy as well as the date of last seizure. The majority of patients were also seizure free, with a date of last seizure of more than 2 years prior to the visit. As foreseen, the majority of patients in our cohort had epilepsy. Generalized tonic-clonic and focal with impairment of consciousness seizure types were the most predominant in our cohort. We concluded that the fields less documented were related to missed ASM, ASM side effects and barriers to adherence. This might be due to later introduction of ASM related fields, compared to the other metrics, which were documented since the beginning of SmartPhrase implementation. When documented, the majority of patients did not miss more than two doses in a row of ASM, did not present side effects to the ASM and did not experience any barriers to adherence. Being documentation of barriers to medication adherence one of the ELHS network-wide interventions, we reported these findings to providers to promote adherence to reporting. Therefore, we conclude that by using a text mining methodology applied to quality data, we can provide meaningful insights and engage neurologists to improve their patient outcomes. Thus, we are able to provide a tool that has the potential to improve the reporting of vital seizure control metrics, thereby improving patient care.

4.2. Prior work

Prior work has been done using QI methodologies applied to epilepsy research. In (Narayanan et al., 2017), the authors describe a methodology for building structured clinical documentation support tools in the EHR to support QI and define best practices in epilepsy. The authors describe how they incorporated these toolkits into their clinical workflow. They also shared the EHR tools developed with other epilepsy clinics as part of a Neurology

Practice Based Research Network and applied these tools to conduct pragmatic trials using subgroup-based adaptive designs. A multidisciplinary QI team conducted analysis of data for prescribed seizure rescue medication doses to identify and improve inappropriately low dose prescriptions (Patel et al., 2020). The QI team identified areas of focus for improvement opportunities and developed the project objective based on the 2017 American Academy of Neurology (AAN) and Child Neurology Society (CNS) Quality Measure. The team created an automated monthly report to monitor prescribed seizure rescue medication dosing compliance. The team successfully decreased provider prescribed and signed under-dosed rescue medication orders by an average of 89 %. In (Cisneros-Franco et al., 2013), based on the AAN quality measures that should be observed at every patient visit, the authors compared the percentage of documentation of each measure before and after the implementation of a new worksheet in a third-level center. Documented measures included seizure type and frequency, etiology, electroencephalogram (EEG), magnetic resonance imaging/ computerized tomography (MRI/CT) head scans, ASM side effects, surgical therapy referral, safety counseling, preconception counseling-and physical exam. The authors concluded that the quality-oriented epilepsy worksheet led to a better practice standardization and documentation of AAN standards for diagnostic and counseling purposes.

In previous literature, methodologies have been presented to obtain seizure control metrics from clinical notes of patients seen at medical centers. A previous study (Barbour et al., 2019) used regexes to identify sudden unexpected death in epilepsy (SUDEP) risk factors in physician notes. Three variables related to SUDEP were considered in the study: generalized tonic-clonic seizure type, refractory epilepsy and potential or previous epilepsy surgery candidacy. Other metrics, such as seizure frequency, acknowledged by the authors as an important risk predictor, were not considered. The authors stated that there were limited NLP tools to reliably evaluate temporal characteristics, and none would be suitable for the analysis. Indeed, extracting key information from unstructured text is challenging. In our case, the SmartPhrase was designed to have specific categories associated with the seizure frequency, thus it was possible to reliable use regexes. The authors (Barbour et al., 2019) concluded that regexes were a feasible method to identify variables related to SUDEP risk and that their methods could be implemented to create large patient cohorts for research. Cui et al. (2012) developed the epilepsy data extraction and annotation (EpiDEA) rule-based system, which extracts information from epilepsy monitoring unit (EMU) discharge summaries. The study also focused on variables of interest for the study of SUDEP: seizure semiology, EEG and MRI patterns, and antiepileptic drug medication. The authors demonstrated the use of EpiDEA for cohort identification through use of an intuitive visual query interface that can be used by clinical researchers. Cui et al. (2014) also developed a rule-based information extraction system called Phenotype Extraction in Epilepsy (PEEP) to automatically identify epilepsy phenotypes with anatomical locations. The variables extracted by PEEP were epileptogenic zone, seizure semiology, lateralising sign, interictal and ictal EEG pattern. The authors demonstrated that their approach was effective in extracting complex epilepsy phenotypes for cohort identification, with an integrated ontology-driven visual query interface. Sullivan et al. (2014) used a machine learning based NLP pipeline to identify a rare epilepsy syndrome from discharge summaries

and EEG reports. The authors concluded that their model could assist physicians to identify the epilepsy syndrome not previously diagnosed, potentially improving seizure control and quality of life for the patients. Fonferko-Shadrach et al. (2019) developed the extraction of epilepsy clinical text (ExECT) system, combining rule-based and statistical techniques to extract information from epilepsy clinic letters. Nine variables were considered: epilepsy diagnosis, epilepsy type, focal seizures, generalized seizures, seizure frequency, medication name (identified when accompanied by a quantity and frequency), CT, MRI and EEG. The authors concluded that their methods could enhance routinely collected data for research and be used in clinical practice to record patient information in a structured manner.

We did not proceed with a comparison between the performance obtained in our study with that of other studies. We developed the NLP approach for semi-structured data while other studies developed methodologies for unstructured text either from clinical letters or discharge summaries, which is more challenging. We concluded that applying our regexes-based approach to semi-structured text enabled us to capture a higher number of seizure control metrics more reliably.

4.3. Limitations

The DI methodology was developed for specific SmartPhrases in two academic medical centers located in the same geographic region (Boston, Massachusetts), both of which use the EPIC EHR and may not be representative of other US and non-US populations limiting the generalizability of our methodology across populations and hospital settings. Since the methodology still assumes knowledge and granted access to the software, future work includes providing any user, neurologists and quality improvers, with an internet-based notebook to easily obtain the preprocessed seizure control information data. Nonetheless, our methodology presents ease of replicability, it is not time costly, and it is free of charge.

4.4. Conclusions

We designed a methodology to improve the quality of reporting key seizure control metrics. We implemented the methodology that captures seizure control information on a weekly basis to be available to neurologists. We concluded that structured data can be feasibly extracted from text notes of epilepsy patients for weekly reporting to a national learning healthcare system. Finally, translating the reported statistics to improvements in standard of care and outcomes for the patients integrated in the ELHS project is needed and proposed as future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Christopher M. McGraw was supported by the National Institutes of Health (NIH-NINDS K08NS118107) and reports no conflict of interest. Lidia M. V. R. Moura was supported by the Centers for Diseases Control and Prevention (U48DP006377), the National Institutes of Health (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 and reports no conflict of interest.

Claire Jacobs received support from Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Advancing Translational Sciences, National Institutes of Health Award UL1 TR002541) and financial contributions from Harvard University and its affiliated academic healthcare centers and reports no conflicts of interest.

M. Brandon Westover was supported by the Glenn Foundation for Medical Research and American Federation for Aging Research (Breakthroughs in Gerontology Grant); American Academy of Sleep Medicine (AASM Foundation Strategic Research Award); Football Players Health Study (FPHS) at Harvard University; Department of Defense through a subcontract from Moberg ICU Solutions, Inc; by the NIH (1R01NS102190, 1R01NS102574, 1R01NS107291, 1RF1AG064312). M. Brandon Westover is a co-founder of Beacon Biosignals and reports no conflict of interest.

Abbreviations:

ASM	Anti-Seizure Medication
BWH	Brigham and Women's Hospital
CDE	Common Data Elements
CRFs	Case Report Forms
DI	Data Integration
EDW	Enterprise Data Warehouse
EHR	Electronic Health Record
ELHS	Epilepsy Learning Healthcare System
ILAE	International League Against Epilepsy
LHS	Learning Healthcare Systems
LineNBR	note line number
MGB	Mass General Brigham
MRN	Medical Record Number
NLP	Natural language processing
NoteID	Note Identification
PatientID	Patient Identification
PROMs	Patient-Reported Outcome Measures
QI	Quality Improvement
Regex	regular expression
SQL	Structured Query Language
SSMS	Microsoft SQL Server Management Studio

References

Bell D, Gachuhi N, Assefi N, 2018. Dynamic clinical algorithms: digital technology can transform health care decision-making. Am. J. Trop. Med. Hyg 98 (1), 9–14. 10.4269/ajtmh.17-0477. [PubMed: 29141738]

Bindman AB, Pronovost PJ, Asch DA, 2018. Funding innovation in a learning health care system. JAMA 319 (2), 119–120. 10.1001/jama.2017.18205. [PubMed: 29270611]

Barbour K, Hesdorffer DC, Tian N, Yozawitz EG, McGoldrick PE, Wolf S, McDonough TL, Nelson A, Loddenkemper T, Basma N, Johnson SB, Grinspan Z, 2019. Automated detection of sudden unexpected death in epilepsy risk factors in electronic medical records using natural language processing. Epilepsia 60 (6), 1209–1220. 10.1111/epi.15966. [PubMed: 31111463]

Budrionis A, Bellika JG, 2016. The learning healthcare system: where are we now? a systematic review. J. Biomed. Inform 64, 87–92. 10.1016/j.jbi.2016.09.018. [PubMed: 27693565]

Cui L, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS, 2012. EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. AMIA Annu Symp Proc. 2012, 1191–1200. [PubMed: 23304396]

Cui L, Sahoo SS, Lhatoo SD, Garg G, Rai P, Bozorgi A, Zhang GQ, 2014. Complex epilepsy phenotype extraction from narrative clinical discharge summaries. J. Biomed Inform 51, 272–279. 10.1016/j.jbi.2014.06.006. [PubMed: 24973735]

Cisneros-Franco JM, Díaz-Torres MA, Rodríguez-Castañeda JB, Martínez-Silva A, Gutierrez-Herrera MA, San-Juan D, 2013. Impact of the implementation of the AAN epilepsy quality measures on the medical records in a university hospital. BMC Neurol. 13, 112. 10.1186/1471-2377-13-112. [PubMed: 23984949]

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 Behav.: EB 117, 107805. 10.1016/j.yebeh.2021.107805.

Elbattah M, Arnaud E, Gignon M, Dequen G, 2021. The role of text analytics in healthcare: a review of recent developments and applications. HEALTHINF 825–832. 10.5220/0010414508250832.

Enticott J, Johnson A, Teede H, 2021. Learning health systems using data to drive healthcare improvement and impact: a systematic review. BMC Health Serv. Res 21 (1), 200. 10.1186/s12913-021-06215-8. [PubMed: 33663508]

Epilepsy Foundation(2020b). What is Epilepsy? Disease or Disorder? https://www.epilepsy.com/learn/about-epilepsy-basics/what-epilepsy (Accessed 11 December 2021).

Epilepsy Foundation, 2020a. Seizure Control https://www.epilepsy.com/aimforzero/seizure-control (Accessed 15 December 2021).

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.

Forsgren L, Beghi E, Oun A, Sillanpää M, 2005. The epidemiology of epilepsy in Europe - a systematic review. Eur. J. Neurol 12 (4), 245–253. 10.1111/j.1468-1331.2004.00992.x. [PubMed: 15804240]

Loring DW, Lowenstein DH, Barbaro NM, Fureman BE, Odenkirchen J, Jacobs MP, Austin JK, Dlugos DJ, French JA, Gaillard WD, Hermann BP, Hesdorffer DC, Roper SN, Van Cott AC, Grinnon S, Stout A, 2011. Common data elements in epilepsy research: development and implementation of the NINDS epilepsy CDE project. Epilepsia 52 (6), 1186–1191. 10.1111/j.1528-1167.2011.03018.x. [PubMed: 21426327]

Margolis PA, Peterson LE, Seid M, 2013. Collaborative chronic care networks (C3Ns) to transform chronic illness care. Pediatrics 131 (Suppl 4), S219–S223. 10.1542/peds.2012-3786J. [PubMed: 23729764]

McLachlan S, Potts HWW, Dube K, Buchanan D, Lean S, Gallagher T, Johnson O, Daley B, Marsh W, Fenton N, 2018. The heimdall framework for supporting characterisation of learning health systems. J. Innov. Health Inform 25 (2), 77–87. 10.14236/jhi.v25i2.996. [PubMed: 30398449]

Menear M, Blanchette M-A, Demers-Payette O, Roy D, 2019. A framework for value-creating learning health systems. Health Res. Policy Syst 17 (1), 79. 10.1186/s12961-019-0477-3. [PubMed: 31399114]

Meskó B, Drobni Z, Bényei É, Gergely B, Győrffy Z, 2017. Digital health is a cultural transformation of traditional healthcare. MHealth 3, 38. 10.21037/mhealth.2017.08.07. [PubMed: 29184890]

Morain SR, Kass NE, Grossmann C, 2017. What allows a health care system to become a learning health care system: results from interviews with health system leaders. Learn. Health Syst 1 (1), e10015 10.1002/lrh2.10015. [PubMed: 31245552]

Narayanan J, Dobrin S, Choi J, Rubin S, Pham A, Patel V, Frigerio R, Maurer D, Gupta P, Link L, Walters S, Wang C, Ji Y, Maraganore DM, 2017. Structured clinical documentation in the electronic medical record to improve quality and to support practice-based research in epilepsy. Epilepsia 58 (1), 68–76. 10.1111/epi.13607.

Ngugi AK, Bottomley C, Kleinschmidt I, Sander JW, Newton CR, 2010. Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. Epilepsia 51 (5), 883–890. 10.1111/j.1528-1167.2009.02481.x. [PubMed: 20067507]

Patel AD, Herbst J, Gibson A, Karn M, Terry D, Debs A, Yarosz S, Parker W, Cohen DM, 2020. Using quality improvement to implement the CNS/AAN quality measure on rescue medication for seizures. Epilepsia 61 (12), 2712–2719. 10.1111/epi.16713. [PubMed: 33063879]

Sarkar D, 2019. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Apress, Bangalore.

Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, Hirsch E, Jain S, Mathern GW, Moshé SL, Nordli DR, Perucca E, Tomson T, Wiebe S, Zhang Y-H, Zuberi SM, 2017. ILAE classification of the epilepsies: position paper of the ILAE commission for classification and terminology. Epilepsia 58 (4), 512–521. 10.1111/epi.13709. [PubMed: 28276062]

Scobie S, Castle-Clarke S, 2019. Implementing learning health systems in the UK NHS: policy actions to improve collaboration and transparency and support innovation and better use of analytics. Learn. Health Syst 4 (1), e10209 10.1002/lrh2.10209. [PubMed: 31989031]

Sullivan R, Yao R, Jarrar R, Buchhalter J, Gonzalez G, 2014. Text Classification towards Detecting Misdiagnosis of an Epilepsy Syndrome in a Pediatric Population. AMIA Annu Symp Proc. 1082–1087. [PubMed: 25954418]

Teede HJ, Johnson A, Buttery J, Jones CA, Boyle DI, Jennings GL, Shaw T, 2019. Australian health research alliance: national priorities in data-driven health care improvement. Med. J. Aust 211 (11), 494–497. 10.5694/mja2.50409. [PubMed: 31733072]

Thimbleby H, 2013. Technology and the future of healthcare. J. Public Health Res 2 (3), e28 10.4081/jphr.2013.e28. [PubMed: 25170499]

World Health Organization (2019a). Epilepsy. https://www.who.int/news-room/fact-sheets/detail/epilepsy (Accessed 29 December 2021).

World Health Organization (2019b). Health Service Data—WHO. https://www.who.int/data/data-collection-tools/health-service-data (Accessed 15 December 2021).

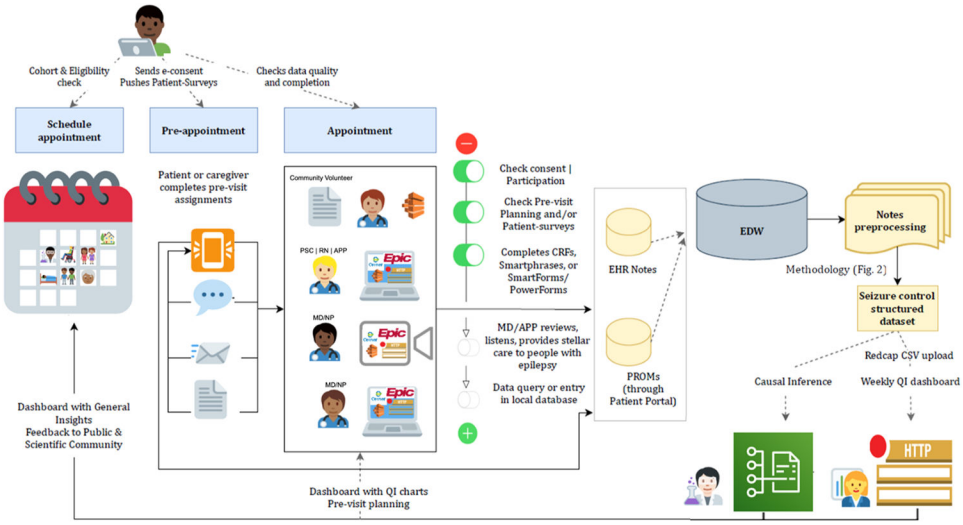

Fig. 1.

Clinical data flow process. EDW – Enterprise Data Warehouse. EHR – Electronic Health Record. QI – Quality Improvement. PROMs – Patient-Reported Outcome Measures.

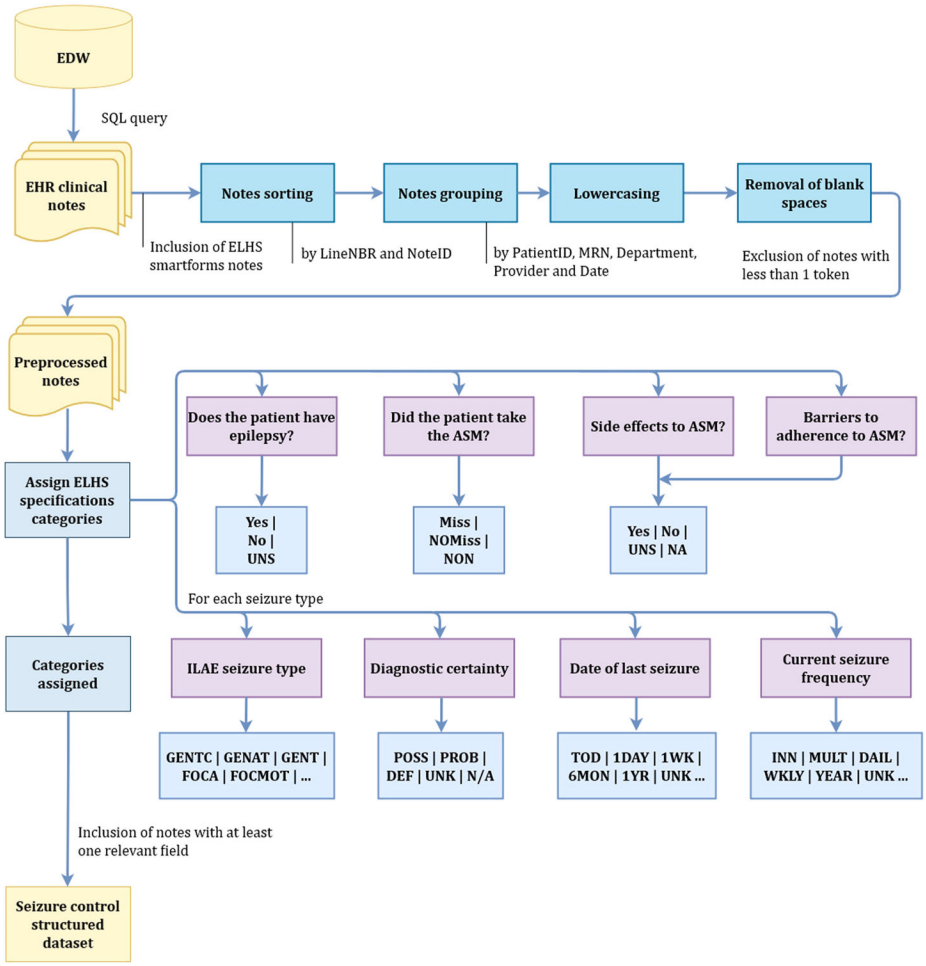

Fig. 2.

Methodology for notes preprocessing. EDW – Enterprise Data Warehouse. EHR – Electronic Health Record. MRN – Medical Record Number. ELHS – The Epilepsy Learning Healthcare System. ILAE – The International League Against Epilepsy. ASM – Anti-seizure medication. ELHS specifications categories and respective acronyms are presented in Supplementary Table 2.

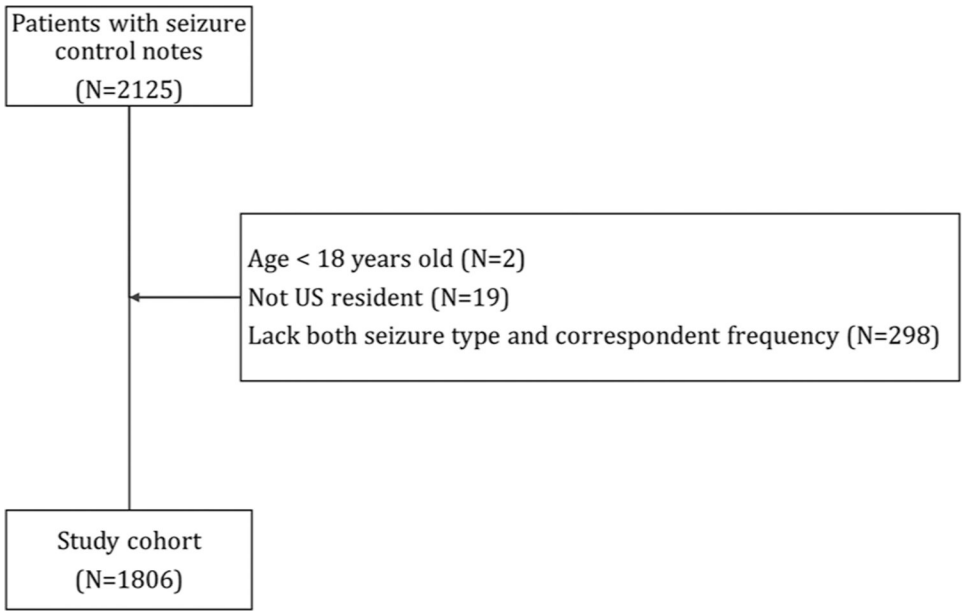

Fig. 3.
Inclusion and exclusion criteria to select the cohort population.

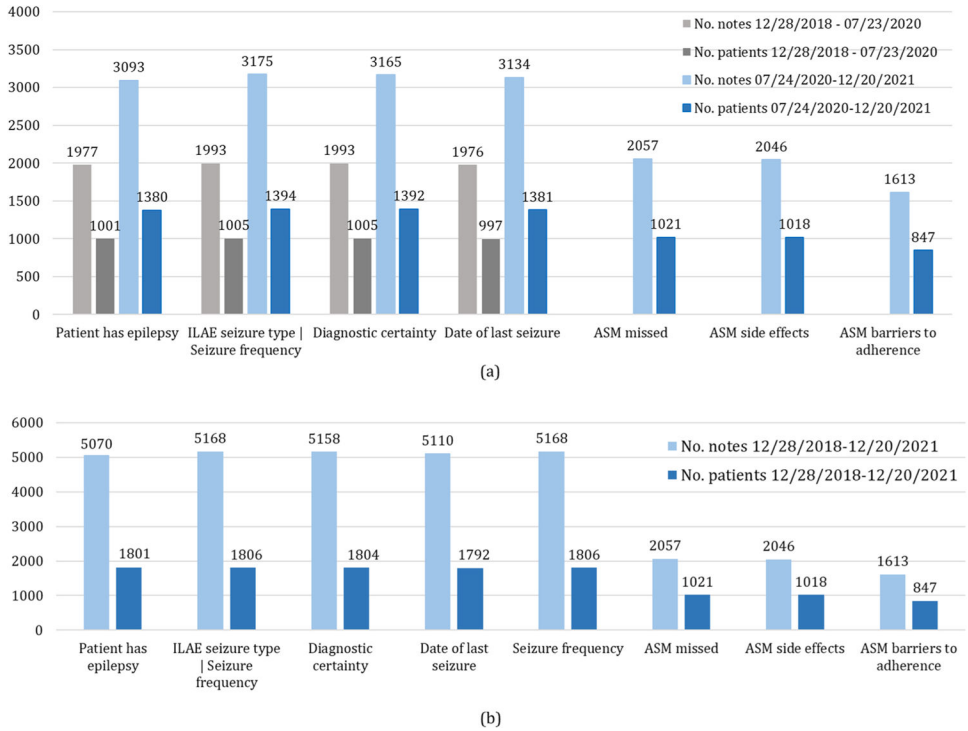

Fig. 4.

Statistics for all notes containing seizure control metrics captured with the regular expressions for the study cohort population. The number of notes and patients are displayed for (a) the period before and after the implementation of ASM documentation, and (b) the study period. ILAE – The International League Against Epilepsy. ASM – Anti-seizure medication.

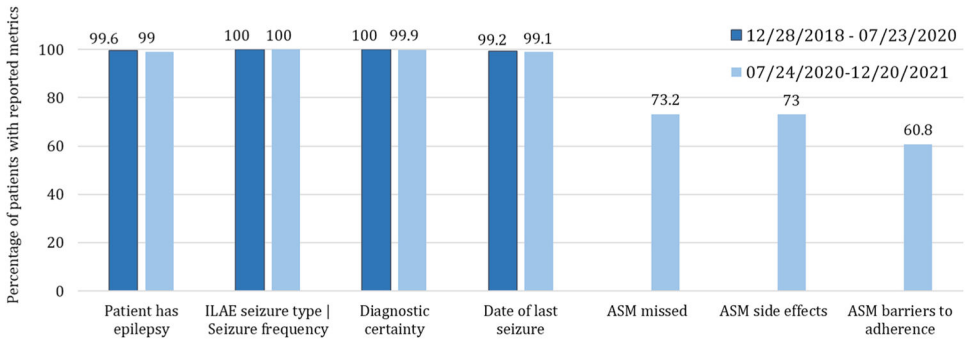

Fig. 5.

Percentage of patients in the study cohort population with documented seizure control metrics before and after ASM documentation date. ILAE – The International League Against Epilepsy. ASM – Anti-seizure medication.

Table 1

Characteristics of the study cohort population.

Characteristic	Study cohort (n = 1806)
Number of clinic visits[(1)]	5168
Age (years, mean (SD))	43.1 (17.0)
Gender, n (%)	
Female	1022 (56.6)
Male	777 (43.0)
Other	*
Race, n (%)	
White	1443 (79.9)
Black or African American	129 (7.1)
Other	234 (13.0)
Ethnicity, n (%)	
Non-Hispanic	1519 (84.1)
Hispanic	135 (7.5)
Other	152 (8.4)
Primary language, n (%)	
English	1696 (93.9)
Spanish	54 (3.0)
Other	56 (3.1)
Marital status, n (%)	
Non-union	1080 (59.8)
Union	700 (38.8)
Other	246 (1.4)
Seizure type at baseline, n (%)	
Focal	1059 (58.6)
Generalized	640 (35.5)
Other	107 (5.9)
Insurance, n (%)	
Private / commercial	818 (45.3)
Medicare / government	636 (35.2)
Other	352 (19.5)

The number of patients is represented by n. SD – standard deviation.

*A number less than 10 is omitted to preserve patient privacy

(1)Number of clinic visits with documented seizure control information including seizure type and frequency.

Table 2

Performance of the regex-based approach algorithm, compared to the gold standard of experts' manual review for 100 clinical notes randomly selected.

Seizure control metric	Number of items (algorithm/experts)	Precision	Recall	F-score
Seizure type				
(a, b, c, d)	100/100, 52/52, 15/16, 2/2	1, 1, 1, 1	1, 1, 0.938, 1	1, 1, 0.968, 1
Diagnostic certainty				
(a, b, c, d)	100/100, 53/53, 18/18, 4/4	1, 1, 1, 1	1, 1, 1, 1	1, 1, 1, 1
Date of last seizure				
(a, b, c, d)	98/99, 50/50, 16/16, 2/2	1, 1, 1, 1	0.989, 1, 1, 1	0.995, 1, 1, 1
Seizure frequency				
(a, b, c, d)	100/100, 49/50, 16/16, 2/2	1, 1, 1, 1	1, 0.980, 1, 1	1, 0.989, 1, 1
Patient has epilepsy	98/98	1	1	1
ASM missed	41/41	1	1	1
ASM side effects	41/41	1	1	1
ASM adherence	30/30	1	1	1
Total	887/890	1	0.996	0.998

ASM – Anti-seizure medication.