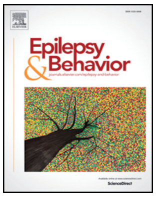
Epilepsy & Behavior 94 (2019) 65–71

Contents lists available at ScienceDirect

Epilepsy & Behavior

journal homepage: www.elsevier.com/locate/yebeh



Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups

Kai He [a], Na Hong [b], Samuel Lapalme-Remis [c], Yangyang Lan [a], Ming Huang [b], Chen Li [a], Lixia Yao [b],*

[a] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shanxi 710049, China
[b] Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA
[c] Division of Neurology, Centre Hospitalier de l'Université de Montréal (CHUM), Montréal, Québec H2X 3E4, Canada

ARTICLE INFO

Article history:
Received 21 November 2018
Revised 1 February 2019
Accepted 1 February 2019
Available online 17 March 2019

Keywords:
Epilepsy treatment
Social media
Text mining
Patient concern

ABSTRACT

Objective: Epilepsy is among the most common chronic neurologic diseases. There is a need for more data on patient perspectives of treatment to guide patient-centered care initiatives. Patients with epilepsy share their experiences on social media anonymously, but little is known about those discussions. Our aim was to learn what patients with epilepsy discuss regarding their condition and identify treatment-related themes from online patient support groups.

Methods: A total of 355,838 posts were collected from three online support groups for patients with epilepsy through a crawling script, and an analytical pipeline was built to identify patient conversation content through leveraging of multiple text mining methods. Results were also displayed by network visualization methods.

Results: Patients with epilepsy sought information about medical treatments, shared their treatment experiences, and sought help from other posters. Key themes related to treatments included the search for optimal personalized treatment strategies as well as identifying and coping with adverse effects.

Significance: This study showed the feasibility of learning about concerns of patients with epilepsy, especially treatment issues, through text mining methods. However, some manual selection and filtering were necessary to ensure quality results for the treatment analysis. Providers should be aware of online discussions and use analyses of such discussions to help guide effective patient engagement during care.

1. Introduction

Epilepsy is among the most common chronic neurologic diseases. The Centers for Disease Control and Prevention (CDC) estimates that prevalence in the United States was approximately 1.2% in 2015 [1]. Epilepsy is a complex disorder, and the nature of the disease and its impact on an individual's social and economic functioning are highly variable, depending on the nature of seizures and their frequency, the patient's social environment, adverse effects of treatments, and medical, cognitive, or psychiatric comorbidities.

Traditionally, epilepsy treatment decisions have been made predominantly by healthcare providers and have been derived and driven by the evidence base and provider experience. However, this approach does not lend itself to incorporating patient needs from a patient perspective. An individualized approach to epilepsy management is needed that incorporates patient preferences, life objectives, and values [2]. The burdens imposed on patients by treatments, which may be complex and difficult to manage, also need to be considered.

Recent research has shown that a patient-centered treatment model promotes greater satisfaction in healthcare [3–5] and that patients should be engaged in their care [6] and participate in decision-making for epilepsy management [7,8]. However, when patient concerns have not been fully investigated and are not fully understood by healthcare teams, patient dissatisfaction and treatment nonadherence may arise. To manage patient care effectively, healthcare providers must engage patients in the decision-making process, and this requires investigation of patient perspectives.

Qualitative methods such as patient interviews and focus groups followed by questionnaires have been generally used to study epilepsy [9] and variety of other diseases such as lung cancer [2] and rheumatic disease [10], from different patient perspectives. Although these methods have enhanced our understanding of patient perspectives, such studies are often expensive and time-consuming to develop. Recruiting a diverse group of participants is often challenging [11–13]. Furthermore, researchers and clinicians often define the focus of the studies, choosing survey questions without understanding patient perspectives [14], and patients may be reluctant to express themselves freely in a formal research setting, leading researchers to miss themes

---

Abbreviations: POS, part of speech; UMLS, Unified Medical Language System; CUI, Concept Unique Identifier.

* Corresponding author at: Department of Health Sciences Research, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA.

E-mail address: yao.lixia@mayo.edu (L. Yao).

https://doi.org/10.1016/j.yebeh.2019.02.002

that are important to patients. Finally, patient preferences, needs, and experiences change with time, culture, and advancement in medicine. For instance, over 100 studies on understanding the information needs and seeking behaviors of pregnant women in the U.S. have been published since late 1980s [15], amounting to a significant cost.

On the other hand, studies have described how social media is now influencing medicine [11–13,16,17]. Patients have increasingly turned to social media platforms, such as online patient support groups, Facebook, and Twitter to share their illness experiences, seek advice and gain support from others with similar conditions, and express views that they do not share with their healthcare providers [18], as they have the opportunity to set the agenda and speak freely on these platforms. An unintended benefit of this activity is that it produces a public record of patient concerns. The increasing number of online patient support groups provides an opportunity for healthcare providers to collect patient opinions and analyze them quantitatively and qualitatively.

Little research, however, has been performed on acquiring and analyzing the opinions and concerns of patients with epilepsy expressed online. This may be due in part to a lack of automated methods and tools for analyzing large amounts of text data on patient forums. Recent studies have shown that text mining is effective for analyzing massive amounts of patient generated text data on health-related issues [19,20].

Therefore, to demonstrate the potential of such methods to learn more about the patient perspectives, we used text mining to quantitatively analyze a large number of online posts generated by patients with epilepsy who posted on online support group sites. Our objective was to identify and analyze the topic areas that patients most discussed on these sites. In addition, further analysis was performed to better understand how posters perceived management of epilepsy and identify key issues important to patients with epilepsy treatment. Exploration of this topic may assist clinicians to better understand their patients, and patient-derived information is essential to designing clinical decision aids [21]. Such research may be complementary to more traditional methods, and could assist in designing qualitative studies and questionnaires.

2. Materials and methods

We collected online posts generated by patients from 3 patient support group sources, where users are able to post anonymously: https://connect.mayoclinic.org/ (332 posts), www.epilepsy.com (109,046 posts), and www.coping-with-epilepsy.com (270,670 posts) (see Table 1). These three sources were selected to reflect different types of message boards: www.epilepsy.com is maintained by the nonprofit Epilepsy Foundation, www.coping-with-epilepsy.com is privately run, and https://connect.mayoclinic.org/ is maintained by a major healthcare provider. We also considered other websites, such as patientslikeme.com and myepilepsyteam.com in the initial study design, but did not include them in final experiments, primarily because of their different privacy policies, where users are required to register with complete or partial disclosure of their identity. However, such a large sample size is still statistically sufficient to provide reliable results [22].

A web crawling script was developed in Python programming language to automatically browse these websites page by page and download all the data from each post, similarly to what most search engines do. To support efficient data analysis, we cleaned the data by removing duplicated posts, including duplications that cite other users occurring in a large discussion thread; short posts with fewer than three letters; short reply posts that quoted another user's comments; and long posts that included only news articles or research papers from somewhere else (and no mention of the user's experience). In total, 355,838 posts were selected from these three sources for data analysis.

An analytical pipeline was built to identify concerns of patients with epilepsy from forum data, and text mining methods were used to extract terms, phrases, and relations and to analyze topics and themes. The key methods included medical concept annotation, frequency analysis, and co-occurrence clustering and visualization (Fig. 1). Each module is further described below.

2.1. Medical concept annotation

To support the semantic-level content analysis, we annotated the medical concepts from the 355,838 filtered posts by using the following five core modules: a tokenizer, a part of speech (POS) tagging module, a chunking module, a MetaMap annotation module, and a manual annotation module. With the tokenizer, 355,838 filtered posts were chunked into a list of sentences, and each sentence was tokenized at the word level. Then, each word was tagged with a POS in the list of sentences by using the Stanford POS tagger [23]. After POS tagging, each processed sentence was regarded as a labeled sequence of words for the next chunking process. For the chunking module, we used the Regular Expression Parser (RegexpParser) in the Natural Language Toolkit (NLTK Project) [24] to identify the noun, adjective, and verb phrases from the labeled sentences. More specifically, the RegexpParser consists of three customized regular expression rules for identifying those phrases in prior to identifying the individual words. If there were no phrases that meet the regular expression rules, the module would return words with noun, adjective, and verb parts of speech. In total, 875,415 different terms were extracted from 355,838 filtered posts. MetaMap [25], a program developed at the National Library of Medicine, was used to identify and normalize biomedical concepts to the Unified Medical Language System (UMLS) metathesaurus [26]. The metathesaurus integrates more than 150 biomedical terminologies and ontologies and contains millions of biomedical and health-related concepts and alternative names. MetaMap helped us to group various names of the same biomedical concepts (such as type 2 diabetes, type II diabetes, diabetes mellitus) and their morphological variations (i.e., singular vs. plural, various tense of verbs) to a UMLS Concept Unique Identifier (CUI). Manual annotation was then performed to verify and select medical concepts from MetaMap annotations that were relevant to the care management experience of patient with epilepsy. For example, the general English terms describing time (i.e., year) or action (i.e., give) were removed, and the concepts with incorrect mapping were also removed through manual verification. The generic and brand medication names were collapsed. Ultimately, 387 medical concepts from 63 meaningful UMLS semantic types, including *sign or symptom*, *disease or syndrome*, and *drugs* were selected for subsequent analyses.

2.2. Frequency analysis

The goal of frequency analysis was to discover concepts that patients with epilepsy mentioned most frequently, especially those related to

Table 1
Description and summary of the three websites used in our study.

Web	Description	Post count	Time span	User count
mayoclinic.org	Ran by a major healthcare provider	332	01/2011–03/2017	104
www.epilepsy.com	Ran by the nonprofit Epilepsy Foundation	109,046	01/2004–05/2017	6356
coping-with-epilepsy.com	Privately run by person with epilepsy and family member	270,670	03/2005–03/2017	34,151

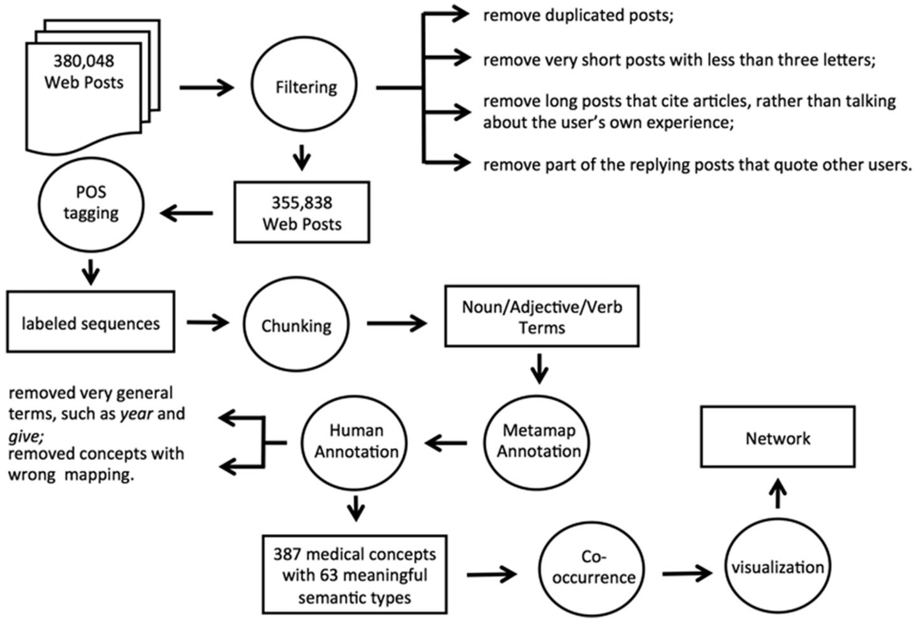


Fig. 1. Text mining pipeline for analyzing data from online epilepsy support groups. POS indicates part of speech.

treatment. The frequency analysis was conducted by counting occurrences of medical concepts and then summing occurrences of medical concepts of the same semantic type as the overall frequency of this semantic type.

2.3. Co-occurrence analyses and visualization

A co-occurrence analysis module was developed to recognize linked relations between each pair of concepts that appeared in the same sentence. Whenever two concepts appeared in the same sentence, a co-occurrence relation of the corresponding medical concept was counted. Using this criterion, a total of 68,570 co-occurrences were found with 387 medical concepts. In response to the large number of co-occurrences, we selected specific treatment semantic types to simplify the co-occurrence networks in order to display the core, important co-occurrence networks. Frequency filters were designed for both the links (co-occurrence relation) and concepts. We used the open source software platform Cytoscape (Cytoscape Consortium) [27] to generate a graphic co-occurrence relation network, which took into consideration semantic type preferences and filter settings. The complete list of the 68,570 co-occurrence pairs of concepts is provided in the Supplemental Table 2.

2.4. Data availability

The datasets used and/or analyzed during this study are available on reasonable request from individuals affiliated with research or healthcare institutions.

3. Results

3.1. What did the patients with epilepsy discuss?

The top 70 high-frequency concepts normalized by UMLS from all posts were identified from the frequency analysis of identified medical concepts (Table 2). Unsurprisingly for message boards dedicated to the concerns of patients with epilepsy, the two most frequent concepts were associated with the condition itself, with the most common concept being “seizures, seizure, fit,” followed by “epileptic, epilepsy.”

Also, ranked near the top of the frequency analysis were concepts relating to healthcare providers, with “doctor, physician” ranked third and “neurologist” ranked fifth. This finding indicated an important concern among board posters with the medical care of epilepsy.

This trend was also evident in the high ranking of concepts related to treatments, with “medication, medicine” (fourth) and “Keppra (levetiracetam)” (sixth) rounding out the top six topics. Given the variety of existing treatments, there were many other high-frequency concepts related to treatment, which, in aggregate, represented an important proportion of the top concepts. Within the top 70 topics, the following had important link to treatments: “Lamictal (lamotrignine)” (17th), “side effect” (19th), “surgery” (21st), “drug” (25th), “Depakote (epilim, valproic acid, valproate, sodium valproate)” (35th), “dose, dosage” (36th), “Dilantin (phenytoin)” (50th), “Tegretol (carbatrol, carbamazepine)” (51st), and “Topamax (topiramate)” (65th). Issues of diagnosis were less highly ranked but still present among the top rankings; “diagnosis, diagnosed” and “electroencephalography (EEG)” were ranked 14th and 16th, respectively.

Other highly ranking concepts were related to communication, personal experiences, and psychosocial concerns, perhaps reflecting the needs of posters sharing their experiences and seeking assistance on message boards. “Feeling, feel, emotion” (seventh), “speak, speaking, talk, talking” (eighth), “life, living” (11th), “read, reading” (13th), “child, kid” (12th), “experience” (15th), and “scare, scared, afraid, fear” (18th) were among those most often used by posters.

An automated analysis of the top 25 semantic types as classified in UMLS is listed in Table 3. Similar to the grouping of high-frequency concepts outlined above, the most commonly cited semantic types referred primarily to epilepsy itself and its consequences (“sign or symptom,” first and “disease or syndrome,” fifth). “Pharmacologic substances,” largely referring to antiseizure medications, were frequently mentioned, ranking second among semantic types. “Mental processes” ranked third, encompassing mostly concepts related to emotions, memory, and communication. Other highly ranking concepts included “family group” (sixth), and “professional or occupational group” (seventh), the latter overwhelmingly referring to physicians.

The raw co-occurrence analysis revealed a dominance of co-occurrences involving the concepts of “seizure” and “epilepsy” with other terms (see Supplemental Table 2). The most common co-occurrence in a sentence was “seizure” and “epilepsy”; the second most common was “seizure” and “medicine.” Among the top co-occurrences, virtually all were pairs linking either “seizure” or “epilepsy” to other concepts, for example, “seizure” and “doctor”

Table 2
Top 70 high-frequency medical concepts.

Concepts 1–35				Concepts 36–70			
No.	Concepts	Count	%	No.	Concepts	Count	%
1	Seizure, seizures, fit	230,526	8.995	36	Dosage, dose	15,100	0.583
2	Epileptic, epilepsy	82,281	3.175	37	Family	14,651	0.565
3	Doctor, physician	56,927	2.196	38	Call	14,452	0.558
4	Medication, medicine	50,761	1.959	39	Husband	14,069	0.543
5	Neurologist	46,766	1.804	40	Checked, check	13,878	0.535
6	Keppra (levetiracetam)	38,568	1.488	41	Falling, fall	13,597	0.525
7	Feeling, feel, emotion	37,239	1.437	42	Bad	13,593	0.524
8	Speak, speaking, talk, talking	36,955	1.426	43	Worse	13,542	0.523
9	Sleeping, sleep	34,830	1.344	44	Write, writing	13,528	0.522
10	Issue, problem	34,690	1.338	45	Symptom	13,405	0.517
11	Life, living	34,212	1.320	46	Care	12,473	0.481
12	Kid, child	33,516	1.293	47	Eating, eat	12,443	0.480
13	Read, reading	29,493	1.138	48	Worry, worried	12,207	0.471
14	Diagnosed, diagnosis	28,094	1.084	49	Wake	12,167	0.469
15	Experience	26,504	1.023	50	Dilantin (phenytoin)	11,864	0.458
16	EEG	25,174	0.971	51	Tegretol (carbatrol, carbamazepine)	11,710	0.452
17	Lamictal (lamotrigine)	24,990	0.964	52	Patient	11,699	0.451
18	Scare, scared, afraid, fear	24,622	0.950	53	MRI	11,498	0.444
19	Side effect	22,310	0.861	54	Stress	11,433	0.441
20	Tonic–clonic seizure, grand mal, grand mal seizure	20,730	0.800	55	Walking, walk	11,366	0.439
21	Surgery	20,718	0.799	56	Turn, turned	11,331	0.437
22	Remember	20,592	0.795	57	Eye	11,251	0.434
23	Understand, understanding	19,991	0.771	58	Result	11,207	0.432
24	Test	19,360	0.747	59	Trigger	10,767	0.415
25	Drug	19,197	0.741	60	Guess	10,711	0.413
26	Body	18,891	0.729	61	Suggest, suggestion	10,637	0.410
27	Hearing, hear	18,249	0.704	62	Edit	10,585	0.408
28	Daughter	18,193	0.702	63	School	10,558	0407
29	Head	17,460	0.674	64	Depression	10,278	0.397
30	Friend	17,410	0.672	65	Topamax (topiramate)	10,229	0.395
31	Show	16,207	0.625	66	Hand	10,164	0.392
32	Hospital	16,164	0.624	67	Mind	9980	0.385
33	Learned, learn, learning	15,646	0.604	68	Partial seizure	9939	0.383
34	Normal	15,408	0.595	69	Memory	9373	0.362
35	Depakote (epilim, valproic acid, valproate, sodium valproate)	15,368	0.593	70	Job	9286	0.360

The complete list can be found in Supplemental Table 1.

(third), "grand mal" and "seizure" (fourth), "epilepsy" and "diagnosis" (sixth), and "seizure" and "Keppra" (14th). Among the top 50 co-occurrences, only five did not contain either the concept of "seizure" or "epilepsy"; these were "medicine" and "doctor" (23rd), "talk" and "doctor" (30th), "neurologist" and "medicine" (44th), "neurologist" and "EEG" (46th), and "magnetic resonance imaging (MRI)" and "EEG" (50th). These findings indicated a central tendency of posters to relate the disease and primary symptom of epilepsy to a wide variety of other concepts, and in particular to treatments.

Table 3
Top 25 high distribution semantic types.

No.	Semantic type	Code	Count	%
1	Sign or symptom	sosy	301,548	11.65
2	Pharmacologic substance	phsu	257,928	9.96
3	Mental process	menp	248,081	9.58
4	Finding	fndg	177,328	6.85
5	Disease or syndrome	dsyn	138,608	5.35
6	Family group	famg	121,622	4.70
7	Professional or occupational group	prog	119,553	4.62
8	Qualitative concept	qlco	108,696	4.20
9	Daily or recreational activity	dora	106,862	4.13
10	Functional concept	ftcn	87,756	3.39
11	Idea or concept	idcn	83,935	3.24
12	Organism function	orgf	81,531	3.15
13	Body part, organ, or organ component	bpoc	62,863	2.43
14	Intellectual product	inpr	41,739	1.61
15	Healthcare activity	hlca	40,133	1.55
16	Diagnostic procedure	diap	39,918	1.54
17	Physiologic function	phsf	39,719	1.53
18	Activity	acty	38,079	1.47
19	Mental or behavioral dysfunction	mobd	37,558	1.45
20	Quantitative concept	qnco	36,622	1.41
21	Manufactured object	mnob	30,085	1.16
22	Food	food	27,406	1.06
23	Body location or region	blor	27,075	1.05
24	Pathologic function	patf	22,310	0.86
25	Population group	popg	22,197	0.86

3.2. Analysis of treatment issues raised by patients

From the top 25 high distribution semantic types, we identified five types that were most closely related to epilepsy treatment ("disease or syndrome," "mental or behavioral dysfunction," "organism function," "pharmacologic substance," and "sign or symptom"). The epilepsy treatment issues were mainly composed of the following topics: subtypes of epilepsy, medications, daily life and the life cycle, emotions and mental health issues, and medical symptoms.

To uncover significant treatment-related concept relations, potential relations between the concepts belong to these five semantic types were derived using co-occurrence network analysis, we chose the top 120 pairs of high-frequency concurrence relations to build a core relations network (Fig. 2). Each node in the co-occurrence network is a medical concept, and a line between two nodes represents the existence of the co-occurrence. The weight of a linking line illustrates the co-occurrence frequency of two concepts, i.e., heavier weight indicates more co-occurrence. The network has 49 nodes and 120 links. The concept "seizure," for example, is from the semantic type "sign or symptom" and is located in the center of the network to denote its frequency. The concepts most related to "seizure" from each of the four other semantic types are "epilepsy," "medicine," "sleep," and "depression."

In addition, the top 50 high-frequency concepts co-occurrence relations were identified to generate a core concepts network, as displayed

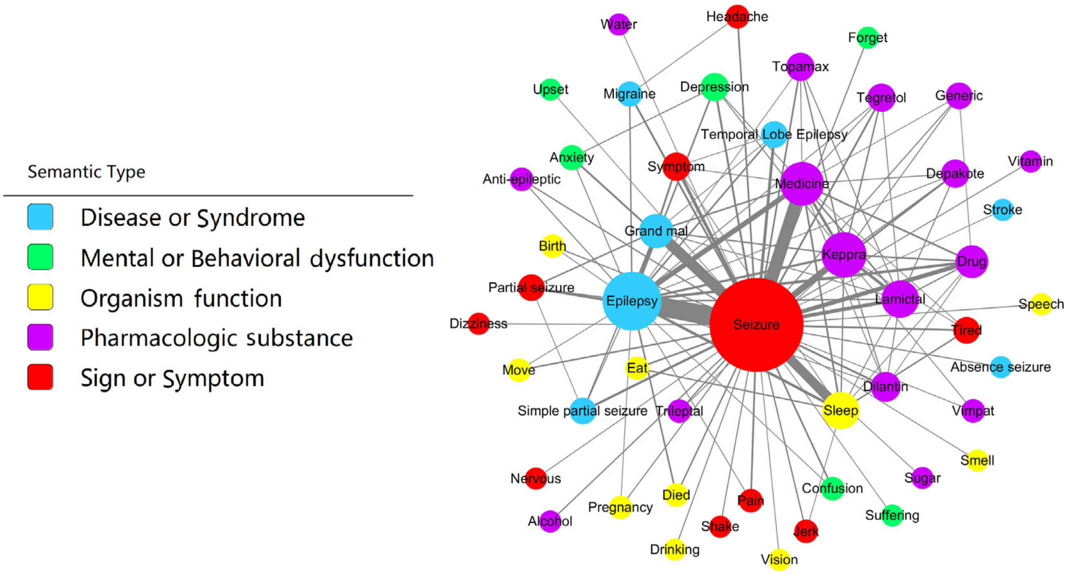


Fig. 2. Core relations network (top 120 high-frequency relations).

in Fig. 3, which clearly revealed the subtopics associated with epilepsy treatment and the cross-relations of these subtopics. Most striking in this network was the clustering relations around the concept of “seizure,” with obviously strong interactions with other concepts. “Seizure” interacted strongly with “epilepsy” (the most important co-occurrence in the dataset), which is at the center of a complex network of interactions with concepts of other semantic types. Concepts co-occurrence relating to medications show them clustering around “seizure” with relatively few direct interactions to signs/symptoms or mental or behavioral dysfunction concepts. “Sleep” is the most common mentioned concept with seizure. Concepts relating to emotions and mental health disturbances show a wider distribution of co-occurrences and appear to modulate the links between epilepsy and many of its characteristics and comorbidities. The overall picture suggests a situation in which the treatment of seizures with medications (usually the main focus of a clinical encounter) is held distinct from the central problem of epilepsy and its symptoms and comorbidities, which are far more linked to emotions and mental health than are seizures.

3.3. Selected concerns regarding treatment

To demonstrate how the posts could provide information on key topics identified in the frequency and co-occurrence analyses above, the 355,838 posts were filtered to seek information of interest specifically relating to posters' perspectives on treatments and were qualitatively reviewed to identify illustrative themes. Posts were identified by searching those containing a co-occurrence between the semantic types “pharmacological substance” and those concepts relating to poster concerns (“questions,” “question,” “wondering,” “wonder,” “want to know,” “why”). Several themes emerged from these treatment-related posts, and are presented here to demonstrate the potential of this method to target specific questions about these topics.

An important theme was a search for strategies regarding for medication management, especially when multiple medications were being taken. A common post theme is that patients enter into highly detailed discussions of their treatments with the apparent goal of establishing an optimal treatment strategy for themselves by recruiting the knowledge and experiences of other posters.

Additionally, medication adverse effects were an important concern, often expressed in the context of struggling to understand what was causing the symptom potentially attributable to their medications.

Another important theme is a concern with the use of generic medications. Posters at times displayed a belief that there was an important therapeutic difference between generic and brand medications (sometimes imparted by healthcare providers), and confusion between different formulations was apparent at times.

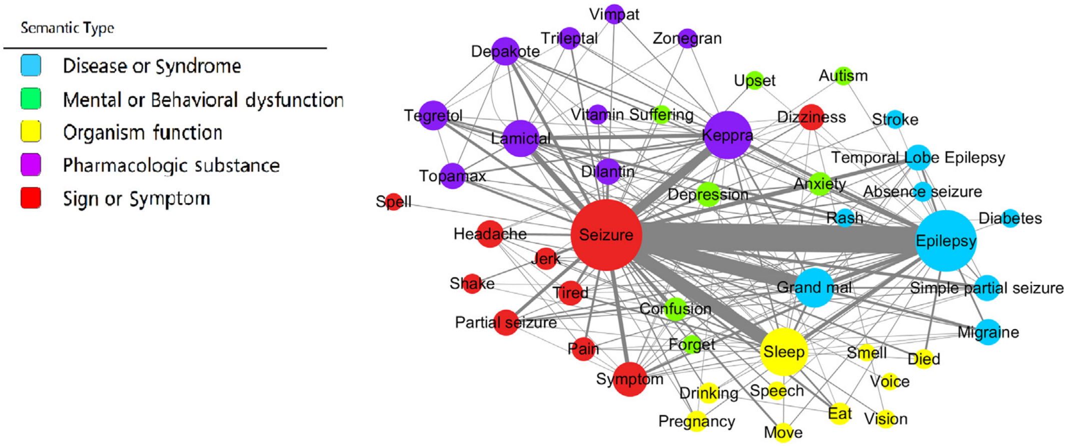


Fig. 3. Core concepts network (top 10 high-frequency concepts in key semantic types).

4. Discussion

This study establishes the feasibility of identifying the interests and concerns of patients with epilepsy by using data from large online patient support groups, although manual selection and filtering of the analysis was necessary to ensure quality results. Such methods can increase our understanding of patient perceptions related to epilepsy and its treatment.

Frequency analysis established that issues relating to treatments and caregivers were among the primary topics discussed by posters and online support groups are a potentially rich source of information on these topics. Posters tended to relate seizures to many other concepts, and in particular, linked medications to seizure control, which was often isolated from other areas of importance to them. In addition, we observed that patients were more inclined to discuss the treatment for their seizure events specifically, and shy away from using the term “epilepsy.” This may indicate that there is still stigma associated with epilepsy, but less so with the term seizures.

Posters often shared their treatment experiences and sought help from other posters; these behaviors suggest that patients were willing to engage in the decision-making process but lacked sufficient knowledge or perceived barriers for engaging providers. In such situations, patients are likely to become less engaged and less compliant with their treatment plan, and they may seek help from other patients. This finding is consistent with results of other related studies, which showed that patients and providers have differing views regarding a patient's understanding of treatments [28].

This study has limitations. First, text mining methods, which can help determine main content from large amounts of free text quickly, may cause some context for the concepts to be lost and may also discriminate against low-frequency concepts. In addition, we noticed that some sentences were improperly tokenized into one sentence. This probably occurred because some sentences were very long, and the patients likely used nonstandard expressions in their online posts, thus decreasing the clarity of the co-occurrence analysis.

Second, we collected data from three online epilepsy support groups, obtaining a greater quantity of data that could be retrieved by traditional quantitative analysis, but by no means were the data complete. Those posting to these sites may not be representative of people posting to all message boards, and patients who post to social media likely have characteristics that make them different from patients who do not. Different concerns may also have been conveyed by patients on sites we did not mine. Consequently, the sample of patients studied is certainly affected by selection bias. Although the findings cannot therefore be considered representative of all patients with epilepsy, they do represent the views of at least a segment of patients with epilepsy, and this information would be useful in the design of further research. Furthermore, although it is likely that the majority of posters are patients with epilepsy, the identity of posters cannot be confirmed, and people who do not have epilepsy, such as the parents of children with epilepsy, were noted to be frequent posters. The findings therefore do not represent only the concerns of people with epilepsy.

Third, these text mining methods are designed to capture the most frequent and common themes from the data. As a result, the findings were generally unsurprising and few previously unanticipated patient concerns regarding epilepsy diagnoses and treatment were identified. These findings may serve as a good validation of our method; but in many novels, unseen themes are no doubt buried in our long list of results. More intelligent computational methods are needed to distinct potentially true and meaningful signals from noises because of poor data quality or computational error. More targeted research with in-depth analysis of relevant posts would also be more effective in identifying novel findings.

5. Conclusion

This study showed the feasibility of using text mining methods to learn concerns of patients with epilepsy, an in particular concerns regarding treatment. A total of 355,838 posts were collected from three online epilepsy support groups, and an analytical pipeline was built to impartially identify the patient concerns. Our method of text mining social media can provide a means for achieving an objective understanding of the issues faced by patients with epilepsy. More targeted analysis seeking to answer specific questions about patient perceptions of treatments is warranted, and could provide evidence for developing patient-centered epilepsy treatment, thus decreasing the gap between patients and providers and encouraging more patient engagement regarding their epilepsy care.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yebeh.2019.02.002.

Author contributions

SLR identified the research questions, interpreted the data results, and drafted portions of the manuscript. KH performed the experiment. NH interpreted the data results. YL and MH collected and cleaned the data. CL and LY designed the entire research project and supervised the study. KH, NH, and LY drafted the manuscript.

Acknowledgment

Funding for this study for Mayo Clinic authors were provided by the Center for Clinical and Translational Science, Mayo Clinic (UL1TR002377) from the National Institutes of Health/National Center for Advancing Translational Sciences and the National Library of Medicine (5K01LM012102).

Funding for this study for Xi'an Jiaotong University authors were provided by The Fundamental Theory and Applications of Big Data with Knowledge Engineering (2016YFB1000903), Project of China Knowledge Centre for Engineering Science and Technology, Innovation team of Ministry of Education of China (IRT-17R86), Innovative Research Group of the National Natural Science Foundation of China (61721002), Ministry of Education-Research Foundation of China Mobile Communication Corp (MCM20160404), and National Science Foundation of China (6177051795).

Conflicts of interest

The authors declared no conflicts of interest.

Ethics approval and consent to participate

In this study, we mined only publicly available information from three websites, without interacting with, intervening, or manipulating/changing the website's environment. The study does not include “human subject” data and is approved by the Office of Research and Compliance without IRB requirement at Mayo Clinic.

We confirm that we have read the journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

References

[1] Zack MM, Kobau R. National and state estimates of the numbers of adults and children with active epilepsy - United States, 2015. MMWR Morb Mortal Wkly Rep 2017;66:821–5.

[2] Chapple A, Ziebland S, McPherson A. Stigma, shame, and blame experienced by patients with lung cancer: qualitative study. BMJ 2004;328:1470.

[3] Charon R. Narrative medicine: honoring the stories of illness. Oxford; New York: Oxford University Press; 2006.

[4] Krumholz HM. Patient-centered medicine: the next phase in health care. Circ Cardiovasc Qual Outcomes 2011;4:374–5.

[5] Laine C, Davidoff F. Patient-centered medicine. A professional evolution. JAMA 1996; 275:152–6.

[6] Bensing J. Bridging the gap. The separate worlds of evidence-based medicine and patient-centered medicine. Patient Educ Couns 2000;39:17–25.

[7] Pickrell WO, Elwyn G, Smith PE. Shared decision-making in epilepsy management. Epilepsy Behav 2015;47:78–82.

[8] Shafer PO. Shared decision-making in epilepsy management–its time has come, but are we missing some concepts? Epilepsy Behav 2015;47:73–4.

[9] Prinjha S, Chapple A, Herxheimer A, McPherson A. Many people with epilepsy want to know more: a qualitative study. Fam Pract 2005;22:435–41.

[10] Haugli L, Strand E, Finset A. How do patients with rheumatic disease experience their relationship with their doctors?: a qualitative study of experiences of stress and support in the doctor–patient relationship. Patient Educ Couns 2004;52: 169–74.

[11] Shepherd A, Sanders C, Doyle M, Shaw J. Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation. BMC Psychiatry 2015;15:29.

[12] Freedman RA, Viswanath K, Vaz-Luis I, Keating NL. Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment. Breast Cancer Res Treat 2016;158:395–405.

[13] Ahmed S, Haines-Saah RJ, Afzal AR, Tam-Tham H, Al Mamun M, Hemmelgarn BR, et al. User-driven conversations about dialysis through Facebook: a qualitative thematic analysis. Nephrol Ther 2017;22:301–7.

[14] Downey L, Au DH, Curtis JR, Engelberg RA. Life-sustaining treatment preferences: matches and mismatches between patients' preferences and clinicians' perceptions. J Pain Symptom Manage 2013;46:9–19.

[15] Sayakhot P, Carolan-Olah M. Internet use by pregnant women seeking pregnancy-related information: a systematic review. BMC Pregnancy Childbirth 2016;16:65.

[16] Savas JA, Huang KE, Tuchayi SM, Feldman SR. Understanding the influence of social media in medicine: lesson learned from Facebook. Dermatol Online J 2014;20.

[17] Beusterien K, Tsay S, Gholizadeh S, Su Y. Real-world experience with colorectal cancer chemotherapies: patient web forum analysis. Ecancermedicalscience 2013; 7:361.

[18] Smith SK, Selig W, Harker M, Roberts JN, Hesterlee S, Leventhal D, et al. Patient engagement practices in clinical research among patient groups, industry, and academia in the United States: a survey. PLoS One 2015;10:e0140232.

[19] Park J, Ryu YU. Online discourse on fibromyalgia: text-mining to identify clinical distinction and patient concerns. Med Sci Monit 2014;20:1858–64.

[20] Raja U, Mitchell T, Day T, Hardin JM. Text mining in healthcare. Applications and opportunities. J Healthc Inf Manag 2008;22:52–6.

[21] Coulter A, Stilwell D, Kryworuchko J, Mullen PD, Ng CJ, van der Weijden T. A systematic development process for patient decision aids. BMC Med Inform Decis Mak 2013;13(Suppl. 2):S2.

[22] van Rijnsoever FJ. (I can't get no) saturation: a simulation and guidelines for sample sizes in qualitative research. PLoS One 2017;12:e0181689.

[23] Stanford Log-linear Part-Of-Speech Tagger; 2019. Retrieved from https://nlp.stanford.edu/software/tagger.shtml.

[24] Bird S, Loper E. NLTK: the natural language toolkit. Proceedings of the ACL 2004 on interactive poster and demonstration sessions. Association for Computational Linguistics; 2004. p. 31.

[25] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17:229–36.

[26] The UMLS Terminology Services (UTS); 2019. Retrieved from https://uts.nlm.nih.gov/home.html.

[27] Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol 2011;696:291–303.

[28] Kelley MS, Jacobs MP, Lowenstein DH, Stewards NEB. The NINDS epilepsy research benchmarks. Epilepsia 2009;50:579–82.