The Application of Text Mining Algorithms In Summarizing Trends in Anti-Epileptic Drug Research

Shatrunjai P. Singh[1], Swagata Karkare[2], Sudhir M. Baswan[3] & Vijendra P. Singh[4]

[1] Lindner College of Business, University of Cincinnati, Ohio, USA

[2] School of Public Health, Boston University, Boston, MA, USA

[3] James L. Winkle College of Pharmacy, University of Cincinnati, Ohio, USA

[4] Department of Internal Medicine, Baptist Easley Hospital, Easley, SC, USA

Correspondence: Shatrunjai P. Singh, Lindner College of Business, University of Cincinnati, Ohio, USA.

Abstract

Content summarization is an important area of research in traditional data mining. The volume of studies published on anti-epileptic drugs (AED) has increased exponentially over the last two decades, making it an important area for the application of text mining based summarization algorithms. In the current study, we use text analytics algorithms to mine and summarize 10,000 PubMed abstracts related to anti-epileptic drugs published within the last 10 years. A Text Frequency – Inverse Document Frequency based filtering was applied to identify drugs with highest frequency of mentions within these abstracts. The US Food and Drug database was scrapped and linked to the results to quantify the most frequently mentioned modes of action and elucidate the pharmaceutical entities marketing these drugs. A sentiment analysis model was created to score the abstracts for sentiment positivity or negativity. Finally, a modified Latent Dirichlet Allocation topic model was generated to extract key topics associated with the most frequently mentioned AEDs. We found the top five most common drugs that appeared from the analysis were Gabapentin, Levetiracetam, Topiramate, Lamotrigine and Acetazolamide. We further listed the key topics associated with these drugs and the overall positive or negative sentiment associated with them. Results of this study provide accurate and data intensive insights on the progress of anti-epileptic drug research.

Keywords: Text Analytics, Anti-Epileptic Drugs, Sentiment Analysis, Topic Modeling

1. Introduction

The unparalleled surge in published biomedical literature has made it difficult to define quantitative and qualitative summarization of a specific topic. Recent advances in computational power have led to an increase in the use of text mining approaches to facilitate the summarization and content review (Khordad & Mercer, 2017; Moradi & Ghadiri, 2017; Zhu et al., 2013). Open source analytical tools can rapidly ingest vast sources and volumes of information which can then be further pipelined into key insights using algorithms like feature extraction, topic modeling and sentiment analysis, allowing accurate summarization (Mishra et al., 2014). These text mining approaches have already been employed in analyzing a wide array of topics like oncology databases (Zhu et al., 2013), impact of financial crises on suicides (Jung et al., 2017), awareness of climate change in rural communities (Bell, 2013) and analyzing the sentiment of diabetes patients on the twitter platform (Salas-Zárate et al., 2017). Although text mining has been employed in several domains of biomedical research, its use remains infrequent in many important therapeutic areas, including neuroscience research (Singh, 2015).

Epilepsy, the fourth most frequent neurological disorder, affects more than sixty million people globally (Singh, 2015; Singh & Karkare, 2017; Trinka et al., 2015; Singh, He, McNamara, & Danzer, 2013; Singh, LaSarge, An, McAuliffe, & Danzer, 2015). The social stigma linked with this condition often primes depression and is frequently associated with a decline in the quality of life (Benson et al., 2016; Hester et al., 2016; Luna et al., 2017). The problem is exacerbated by the confusion of focusing research efforts on multiple anti-epileptic drugs (AEDs), some of which show mixed results in the refractory epileptic populations (Ahmad et al., 2017; de Biase, Valente, Gigli, & Merlino, 2017; Nolan, Marson, Weston, & Tudur Smith, 2015; Pellock et al., 2017; Turner & Perry, 2017). The sheer volume of new research on AEDs cripples any meaningful insight generation.

In this study, we analyze 10,000 PubMed abstracts related to AEDs with the end goal of content summarization and

insight generation. Abstracts containing US FDA curated list of drugs were identified and analyzed for drug frequency, mode of action and the pharmaceutical entities manufacturing the most frequent drugs were acknowledged. A modified latent dirichlet analysis algorithm with a bigram tokenizer (mLDA) was used to extract key topics discussed in these abstracts. Finally, sentiment analysis was utilized to analyze which of these anti-epileptic drugs are promising candidates for further research based on associated positive sentiments.

2. Methods

2.1 Data Collection

We used an R-software based PubMed scrapper to download 10,000 abstracts positive either of these keywords: 'anti-epileptic drugs', 'anti-convulsant drugs' and 'AED'. Only abstracts published between 01/01/2007 to 01/01/2017 were included in the study. Papers with no abstracts or written in languages other than English were filtered out. The raw abstract data was uploaded to a public repository for open access (Singh & Karkare, 2018) . The raw R code used for the analysis was deposited in a *Github* repository (https://github.com/shatrunjai/aed_pubmed). A document term matrix (DTM) was created from the abstracts and was compared to the list of drugs approved in the last decade, obtained from the US Food and Drug Administration website (https://www.fda.gov/Drugs.htm). Only abstracts focusing on at least one of these drugs was included for further analysis.

2.2 Data Processing

Collected abstracts were scrubbed for numbers, non-English characters and stop words. The Stanford stop words list was used as the default stop word repository (https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms

-stop-words-1.html). Stemming of abstract was conducted according to the Porter stemmer (Porter, 1980). A document term matrix was created as described in Stanford NLP (https://nlp.stanford.edu/).A Term-Frequency-Inverse Document Frequency (TF-IDF) matrix was created and further frequency calculations were performed only on relevant TF-IDF terms as described in (Jones, 1972).The frequency matrix had a mean word frequency of 272 words and a standard deviation of 17 words. Words with frequency cut off two standard deviations from the mean word frequency were filtered from the list.

2.3 Modified LDA Based Topic Modelling

Latent Dirichlet Allocation (LDA) is a well-defined, unsupervised, generative, probabilistic method for modeling data and is frequently used in topic modeling (Blei, Ng, & Jordan, 2003). We created a modified Latent Dirichlet Allocation (mLDA) algorithm which assumes that each document can be denoted as a probabilistic distribution over latent topics and that the topic distribution in all documents share a common Dirichlet prior distribution. We also included a bigram tokenizer to better represent scientific abstracts. Each latent topic in the mLDA model is also represented as a probabilistic model over words and the word distributions of topics share a common Dirichlet prior distribution as well. Given a corpus M consisting of N documents, with document d having K_d words ($d \in \{1,..., N\}$), mLDA models M according to the following generative process (Blei et al., 2003; Li et al., 2016):

(a) Select a multinomial distribution φ_t for topic t ($t \in \{1,..., T\}$) from a Dirichlet prior distribution with parameter β.

(b) Select a multinomial distribution θ_d for document d ($d \in \{1,..., N\}$) from a Dirichlet prior distribution with parameter α.

(c) For a word w_n ($n \in \{1,..., K_d\}$) in document d,

(i) Select a topic z_n from θ_d.

(ii) Select a word w_n from φ_{zn}.

This generative process has words in documents are the only detected variables whereas others are latent variables (φ and θ) and hyper parameters (α and β). In order to deduce the latent variables and hyper parameters, the probability of experiential data M is calculated as follows:

$$p(M|\alpha, \beta) = \prod d = 1N \int p(\theta d|\alpha)\left(\sum n = 1Kdp(zdn|\theta d)p(wdn|zdn, \varphi)P(\varphi|\beta)\right)d\theta dd\varphi$$

Due to the coupling between θ and φ in the integrand (above equation), the precise implication in mLDA is obstinate (Blei et al., 2003).The number of topics was selected according to the Rate of Perplexity Change (RPC) previously described by Zhao and colleagues (Zhao et al., 2015). This algorithm yielded two key topics on average which were further curated manually.

2.4 Sentiment Analysis

To evaluate sentiment for each abstract, the *Sentiment Analysis* and *Tm* libraries were used within R-software ,Version 0.98.109 ("Text Mining Infrastructure in R | Feinerer | Journal of Statistical Software," n.d.). *Sentiment Analysis* algorithm is a well-established sentiment analysis (SA) protocol and has been cited by over a 1000 journal publications according to google scholar. *Sentiment Analysis* and *Tm* packages assign three sentiment scores ("positive," "negative," and "neutral") to each word, based on a generalized classification system developed by the authors which uses a combination of human-annotated and Artificial Intelligence based sentiment scoring algorithms (Bagheri & Islam, 2017). Further, we employed the "bag-of-words" approach which has been established to be very dependable for document-level SA, with aggregate-level performance approximately equivalent to more refined methods (Gayle & Shimaoka, 2017).

For the current study, nouns were excluded from the analysis as they contain little to no information (Pinheiro, Prado, Ferneda, & Ladeira, 2015). The sentiment of each abstract was calculated by combining the scores of all pertinent word tokens. A sentiment score ranging from -1 to $+1$ was allocated for each abstract based on the assessed grade of negative or positive sentiment. For further analysis and visualization, unstandardized scores were normalized to a distribution with a mean of zero ($\bar{x}=0$) and standard deviation of one ($\sigma_{\bar{x}}=1$). All abstracts were assigned values of 'positive' (score>+1),'negative' (score>-1) or 'neutral' (-1<score<+1).

2.5 Machine Learning

The *Sentiment Analysis* package uses a one class support vector machine (SVM) algorithm to classify the expressions and phrases within the abstracts based on Stanford core NLP trained algorithm. SVM is a supervised analytical method that classifies based on the degree to which the several input cases (i.e., expression vectors) predict a given binary class, like the presence of absence of positive sentiment (Salas-Zárate et al., 2017). All input terms, i.e. the bigrams can thus be assessed in terms of "importance" with respect to a given label (Gayle & Shimaoka, 2017).The classifier was retrained on a 7000-abstract sample curated dataset optimized for misclassification rate, precision and recall metrics.

2.6 Statistical Analyses

Microsoft SQL Server (version 2012) was used to query the dataset for different clone compositions, and statistical analysis was performed using R-statistical software (Version 0.98.109). Significance was determined using a two-tailed Student's t-test for data that met assumptions of normality and equal variance. The Mann-Whitney rank sum test was used for non-normal data. Proportions were compared using z-tests. Values presented are means $\pm$ SEM or medians [range], as appropriate. The experiment-wise error was conservatively set at 0.001 (Cumming, 2010). Corrections for multiple comparisons were done using a Bonferroni correction.

2.7 Figure Preparation

The results from R-software were exported into csv files which were imported into Tableau (version 8.0) or Microsoft Excel (version 2013) which were then used to create graphs and visualizations. Tables were created in Microsoft Word (version 2013).

3. Results

3.1 Characterizing the Most Published Anti-Epileptic Drugs in the Last 10 Years

To study AEDs that appeared in PubMed abstracts (2007-2017), an R scrapper was used to parse 10,000 PubMed abstracts. To identify abstracts specifically related to AEDs, this scrapped dataset was cross-referenced with the United States Food and National Drug database (US FDA) of drugs. A total of 130 drugs (Figure 1) with a mean of 69.34 abstracts per drugs and a standard deviation of 22.03 abstracts per drugs were identified. The top 5 most frequent drugs were: Gabapentin (abstract count=1371, Figure 1), Levetiracetam (abstract count=1304, Figure 1), Topiramate (abstract count=1027, Figure 1), Lamotrigine (abstract count=989, Figure 1) and Acetazolamide (abstract count=518, Figure 1). A year-by-year frequency of selected drug abstracts was performed for all the drugs beginning the year 1980 (Figure 2) to follow their research trends.

3.2 Characterizing Drug Class of the most Published Drugs

For all drugs, their pharmaceutical drug categorization was evaluated by using FDA definitions (https://www.fda.gov/drugs/informationondrugs/ucm079436.htm). As expected, Anti-Epileptic agents and CNS activity suppression agents were at the top of the list of our drug matches (Figure 3). However, cox-2 inhibitors, mood stabilizers, cytochrome p450-2C19 inhibitors, analgesics and serotonin reuptake inhibitors also frequent in the class of researched AEDs (Figures 3), reflecting the diversity in research initiatives.

3.3 Characterizing Pharmaceutical Industries with the Most Published Drugs

Next, the pharmaceutical companies associated with the highest frequency of drug mentions in the 10,000 abstracts selected for the study were extracted (Figure 4). Some companies had more than 5 drugs (Sagent Pharmaceuticals and Zydus Pharmaceuticals Inc. with 14 and 9 drugs, respectively). Zydus Pharmaceutical's Topiramate along with the other 8 drugs appears to lead the list in terms of the number of drugs and the frequency of abstract mentions. However, other companies like A-S Medication Solutions which despite having only one drug (Gabapentin), were still top-ranked in abstract mention frequency (Figure 4).

3.4 Using Sentiment Analysis to Score the Abstracts with the Top Anti-Epileptic Drugs

A sentiment analysis was performed on all the abstracts containing the keyword 'anti-epileptic drugs' or 'AED' or 'anti-convulsion drugs'. An initial analysis revealed a strong correlation between negative sentiment and the frequency of abstract mention (Table 1, correlation coefficient=0.68). To correct for this, a normalized sentiment score (Sentiment-Sentiment $_{mean}$/Sentiment $_{S.D.}$) was calculated for each drug (Table 1). The sentiment value/abstract correlation was manually tested for accuracy. Drugs Lisinopril (normalized sentiment score= -3.0) and Telmisartan (normalized sentiment score= 3.0) had the highest positive normalized sentiment of all the drugs, indicating that these appeared in abstracts with positive connotations ('positive outcome', 'no side effects') more often than other drugs. Conversely, Ethosuximide (normalized sentiment score= -0.9) and Meloxicam (normalized sentiment score= -2.3) had the most negative sentiment, indicating appearance in abstracts with negative connotations ('negative outcomes', 'side effects').

3.5 A modified Latent Dirichlet Algorithm Reveals Topics Associated with The Top 5 Most Mentioned Anti-Epileptic Drugs

An mLDA algorithm was employed to identify the key topics being discussed in the papers associated with the top 5 drug mentions (Table 2). Key words associated with the top topic indicated research on the lines of spinal surgery and pain outcomes. Levetiracetam was associated with topics including its use in refractory and generalized seizure, response bias by gender and its association with Brivaracetam. Topiramate was associated with topics including long term side effects, the development of drug-resistance, and its effect on Lennox-Gastaut syndrome. Acetazolamide was associated with one topic indicating research on its effect on visual acuity and macular degeneration. Finally, Lamotrigine was associated with one topic indicating possible side effects of dry mouth and blood spots at higher concentration of the drug.

4. Discussion

In this study, we use text analytics algorithms to summarize the latest development in anti-epileptic drug research. We mined the top five drugs that have been extensively published in PubMed, elucidate the pharmaceutical entities manufacturing/marketing these drugs, and also provided sentiment based direction on how this research is trending. Finally, we created an mLDA based topic modelling algorithm to discuss key topics associated with these drugs.

The most popular AED's conventionally used as first line treatment include primidone, ethosuximide, benzodiazepines, carbamazepine and phenobarbital. In the last 20 years, the Food and Drug Administration (FDA) has further approved twelve new AED's and have a longer list of these drugs in the clinical trial pipelines (Asconapé, 2010). Although all of these compounds have been used to treat epilepsy for more than a century a true anti-epileptic drug effective against all seizure grades and all demographics is still unavailable and approximately 30% of patients with epilepsy do not respond to any existing AEDs (Glauser et al., 2006; Singh, 2015). This has fueled basic research into new pharmacological agents with better safety and tolerability, ease of use and better titration rate, fewer potential interactions, and increased efficacy in comorbidities (Azar & Abou-Khalil, 2008). The resultant research from studies on different aspects of multiple AEDs has often made research summarization difficult and calls for newer computational approaches.

PubMed, the most extensively used warehouse of biomedical literature comprises of more than 20 million abstracts and is increasing at a frequency of over 90,000 abstracts per year: the quantity of articles added each year to PubMed has increased three times in the last 10 years (Andronis, Sharma, Virvilis, Deftereos, & Persidis, 2011). As research on a solitary subject may extend across numerous scientific areas and technical journals, it is progressively problematic for scientists to trail all advances in their area of work. The dispersal of information to many different journals and scientific subgroups has created and 'islets of scientific knowledge' and has led to the improvement of literature mining approaches pointing to link ideas and opinions that are not cited in the same editorial. The process of deducing implied knowledge from apparently unrelated concepts has been named literature-based discovery (LBD) (Andronis et al., 2011). These LBD methods have been used in the past for the purpose of theory ideation in association with drug discovery. Some of these LBD techniques include PubMed text mining, TF-IDF based keyword generation, unsupervised document clustering, literature modelling, sentiment analysis and topic modelling techniques. In the current study, we use a subset of these techniques for AED centered research summarization.

The most frequently studied AED was found to be Gabapentin, which is indicated for the treatment of postoperative

neuralgia in adults and for treating partial onset seizures in both pediatric and adult patients (Goa & Sorkin, 1993). Although the exact mode of Gabapentin action is unknown, it has been suggested that its activity depends on its interaction with voltage-gated calcium channels (Goa & Sorkin, 1993). Interestingly, topic modeling revealed the keywords 'pain' and 'spinal surgery' to be associated with this drug. However, although gabapentin is commonly used in pain management, its use in post-operative pain and spinal surgery is controversial (Chang, Challa, Shah, & Eloy, 2014; Singh, Singh, Fatima, Kubo, & Singh, 2008; Yu, Ran, Li, & Shi, 2013).

Levetiracetam, the second most commonly researched anti-epileptic drug, is indicated as an adjunctive therapy in the treatment of partial onset seizures in patients ≥16 years of age with epilepsy(Deshpande & Delorenzo, 2014; Zheng, Du, & Wang, 2015). The precise mechanism(s) by which Levetiracetam exerts its antiepileptic effect is unknown, but studies suggest that this agent acts as a neuromodulator and treats seizures by inhibiting presynaptic calcium channels (Deshpande & Delorenzo, 2014). Topic modeling from this study revealed recent efforts towards comparing the efficacy of Levetiracetam to Brivaricetam which has been a topic of increasing interest over the year (Crepeau & Treiman, 2010; Lyseng-Williamson, 2011).

Topiramate is used as a monotherapy in children of ages two and above and as an adjunctive therapy for adults. Its use is children is specifically indicated for seizures related with Lennox-Gastaut syndrome (LGS) (Crumrine, 2011; Donegan, Dixon, Hemming, Tudur-Smith, & Marson, 2015; Hoy, 2016). Topic modeling showed a strong association of this agent with Lennox-Gastaut syndrome, a disorder which initiates seizures in children (Crumrine, 2011; Singh, 2016; Singh et al., 2016; VanStraten & Ng, 2012).

Acetazolamide, a carbonic anhydrase inhibitor is indicated for the treatment of centrencephalic epilepsies (petit mal, unlocalized seizures) and is also a popular drug for the treatment of glaucoma (Reiss & Oles, 1996; Millichap & Aymat, 1967). Results of the topic modeling used in this study support a strong association of this drug with keywords like 'macular', 'visual', 'acuity', all of which are glaucoma-related terms referring to the discovery of its anti-epileptic properties during treatment of glaucoma patients (Lyall, 2008). Lamotrigine is an antiepileptic drug indicated as an adjunctive therapy in children above the ages of two specifically for primary generalized tonic-clonic seizures. It is also indicated for the treatment of bipolar disorder in patients (Ramaratnam, Panebianco, & Marson, 2016). Although the mechanism of action of this drug is unknown, in vitro pharmacological studies suggest that lamotrigine inhibits voltage-sensitive sodium channels, thereby stabilizing neuronal membranes and consequently modulating presynaptic transmitter release of excitatory amino acids (e.g., glutamate and aspartate). Topic modeling revealed the association of this drug with the terms 'dried blood spots', which suggests that research efforts have been focused on evaluating the safety profile of this drug, specifically in causing blood dyscrasias (Krasowski & McMillin, 2014; Milosheska, Grabnar, & Vovk, 2015; Baswan, Li, LaCount, & Kasting, 2016; Singh et al., 2016; Singh & Singh, 2017).

Sentiment analysis suggests that despite these drugs being well-established and approved lines of therapy in the treatment of a variety of epilepsies and seizures, all 5 drugs were associated with a negative sentiment. This indicates the possibility of mixed results in at least a subset of these research studies. Further, these findings suggest potential unmet need in the area of epilepsy treatment due to the dearth of positive sentiments surrounding these pharmacological agents.

5. Conclusion

This study demonstrates that although research efforts surrounding anti-epileptic treatments are moving in the right direction, there is an unmet need when it comes to the associated sentiments of researchers towards the most frequently studied agents. Despite the potential utility of these drugs in the treatment of epilepsy, their use in treatment could be hindered due to associated negative sentiments. Even though this study delineates the key topics surrounding AED research in the last decade, further research efforts should be conducted to understand the causal relationship between the negative sentiments and the pharmacological profile of these agents. Understanding these causative efforts can help lead the way for pharmaceutical manufacturers to devote research efforts towards improving the profiles of their drugs to better suit the needs of the patients.

References

Ahmad, K. A., Desai, S. J., Bennett, M. M., Ahmad, S. F., Ng, Y.-T., Clark, R. H., & Tolia, V. N. (2017). Changing antiepileptic drug use for seizures in US neonatal intensive care units from 2005 to 2014. *Journal of Perinatology: Official Journal of the California Perinatal Association*, *37*(3), 296–300. https://doi.org/10.1038/jp.2016.206

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., & Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, *12*(4), 357–368. https://doi.org/10.1093/bib/bbr005

Asconapé, J. J. (2010). The selection of antiepileptic drugs for the treatment of epilepsy in children and adults. *Neurologic Clinics*, *28*(4), 843–852. https://doi.org/10.1016/j.ncl.2010.03.026

Azar, N. J., & Abou-Khalil, B. W. (2008). Considerations in the choice of an antiepileptic drug in the treatment of epilepsy.

Seminars in Neurology, *28*(3), 305–316. https://doi.org/10.1055/s-2008-1079335

Bagheri, H., & Islam, M. J. (2017). Sentiment analysis of twitter data. *ArXiv:1711.10377 [Cs]*. Retrieved from http://arxiv.org/abs/1711.10377

Baswan, S. M., Li, S. K., LaCount, T. D., & Kasting, G. B. (2016). Size and Charge Dependence of Ion Transport in Human Nail Plate. *Journal of Pharmaceutical Sciences*, *105*(3), 1201–1208. https://doi.org/10.1016/j.xphs.2015.12.011

Bell, E. J. (2013). Climate change and health research: has it served rural communities? *Rural and Remote Health*, *13*(1), 2343.

Benson, A., O'Toole, S., Lambert, V., Gallagher, P., Shahwan, A., & Austin, J. K. (2016). The stigma experiences and perceptions of families living with epilepsy: Implications for epilepsy-related communication within and external to the family unit. *Patient Education and Counseling*, *99*(9), 1473–1481. https://doi.org/10.1016/j.pec.2016.06.009

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Chang, C. Y., Challa, C. K., Shah, J., & Eloy, J. D. (2014). Gabapentin in acute postoperative pain management. *BioMed Research International*, *2014*, 631756. https://doi.org/10.1155/2014/631756

Crepeau, A. Z., & Treiman, D. M. (2010). Levetiracetam: a comprehensive review. *Expert Review of Neurotherapeutics*, *10*(2), 159–171. https://doi.org/10.1586/ern.10.5

Crumrine, P. K. (2011). Management of seizures in Lennox-Gastaut syndrome. *Paediatric Drugs*, *13*(2), 107–118. https://doi.org/10.2165/11536940-000000000-00000

Cumming, G. (2010). Replication, p rep, and confidence intervals: comment prompted by Iverson, Wagenmakers, and Lee (2010); Lecoutre, Lecoutre, and Poitevineau (2010); and Maraun and Gabriel (2010). *Psychological Methods*, *15*(2), 192–198. https://doi.org/10.1037/a0019521

de Biase, S., Valente, M., Gigli, G. L., & Merlino, G. (2017). Pharmacokinetic drug evaluation of lacosamide for the treatment of partial-onset seizures. *Expert Opinion on Drug Metabolism & Toxicology*, *13*(9), 997–1005. https://doi.org/10.1080/17425255.2017.1360278

Deshpande, L. S., & Delorenzo, R. J. (2014). Mechanisms of levetiracetam in the control of status epilepticus and epilepsy. *Frontiers in Neurology*, *5*, 11. https://doi.org/10.3389/fneur.2014.00011

Donegan, S., Dixon, P., Hemming, K., Tudur-Smith, C., & Marson, A. (2015). A systematic review of placebo-controlled trials of topiramate: How useful is a multiple-indications review for evaluating the adverse events of an antiepileptic drug? *Epilepsia*, *56*(12), 1910–1920. https://doi.org/10.1111/epi.13209

Gayle, A., & Shimaoka, M. (2017). Public Response to Scientific Misconduct: Assessing Changes in Public Sentiment Toward the Stimulus-Triggered Acquisition of Pluripotency (STAP) Cell Case via Twitter. *JMIR Public Health and Surveillance*, *3*(2). https://doi.org/10.2196/publichealth.5980

Glauser, T., Ben-Menachem, E., Bourgeois, B., Cnaan, A., Chadwick, D., Guerreiro, C., … Tomson, T. (2006). ILAE treatment guidelines: evidence-based analysis of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes. *Epilepsia*, *47*(7), 1094–1120. https://doi.org/10.1111/j.1528-1167.2006.00585.x

Goa, K. L., & Sorkin, E. M. (1993). Gabapentin. A review of its pharmacological properties and clinical potential in epilepsy. *Drugs*, *46*(3), 409–427. https://doi.org/10.2165/00003495-199346030-00007

Hester, M. S., Hosford, B. E., Santos, V. R., Singh, S. P., Rolle, I. J., LaSarge, C. L., … Danzer, S. C. (2016). Impact of rapamycin on status epilepticus induced hippocampal pathology and weight gain. *Experimental Neurology*, *280*, 1–12. https://doi.org/10.1016/j.expneurol.2016.03.015

Hoy, S. M. (2016). Topiramate Extended Release: A Review in Epilepsy. *CNS Drugs*, *30*(6), 559–566. https://doi.org/10.1007/s40263-016-0344-5

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Khordad, M., & Mercer, R. E. (2017). Identifying genotype-phenotype relationships in biomedical text. *Journal of Biomedical Semantics*, *8*(1), 57. https://doi.org/10.1186/s13326-017-0163-8

Krasowski, M. D., & McMillin, G. A. (2014). Advances in anti-epileptic drug testing. *Clinica Chimica Acta; International Journal of Clinical Chemistry*, *436*, 224–236. https://doi.org/10.1016/j.cca.2014.06.002

Li, D., Wang, Z., Wang, L., Sohn, S., Shen, F., Murad, M. H., & Liu, H. (2016). A Text-Mining Framework for Supporting Systematic Reviews. *American Journal of Information Management*, *1*(1), 1–9.

Luna, J., Nizard, M., Becker, D., Gerard, D., Cruz, A., Ratsimbazafy, V., … Preux, P.-M. (2017). Epilepsy-associated levels of perceived stigma, their associations with treatment, and related factors: A cross-sectional study in urban and rural areas in Ecuador. *Epilepsy & Behavior: E&B*, *68*, 71–77. https://doi.org/10.1016/j.yebeh.2016.12.026

Lyall, D. A. M. (2008). Unexpected control of a patient's refractory epilepsy when treating glaucoma with acetazolamide. *Canadian Journal of Ophthalmology. Journal Canadien D'ophtalmologie*, *43*(3), 377. https://doi.org/10.3129/i08-036

Lyseng-Williamson, K. A. (2011). Levetiracetam: a review of its use in epilepsy. *Drugs*, *71*(4), 489–514. https://doi.org/10.2165/11204490-000000000-00000

Millichap, J. G., & Aymat, F. (1967). Treatment and prognosis of petit mal epilepsy. *Pediatric Clinics of North America*, *14*(4), 905–920.

Milosheska, D., Grabnar, I., & Vovk, T. (2015). Dried blood spots for monitoring and individualization of antiepileptic drug treatment. *European Journal of Pharmaceutical Sciences: Official Journal of the European Federation for Pharmaceutical Sciences*, *75*, 25–39. https://doi.org/10.1016/j.ejps.2015.04.008

Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S., Mostafa, J., & Del Fiol, G. (2014). Text summarization in the biomedical domain: a systematic review of recent research. *Journal of Biomedical Informatics*, *52*, 457–467. https://doi.org/10.1016/j.jbi.2014.06.009

Moradi, M., & Ghadiri, N. (2017). Different approaches for identifying important concepts in probabilistic biomedical text summarization. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/j.artmed.2017.11.004

Nolan, S. J., Marson, A. G., Weston, J., & Tudur Smith, C. (2015). Carbamazepine versus phenobarbitone monotherapy for epilepsy: an individual participant data review. *The Cochrane Database of Systematic Reviews*, (7), CD001904. https://doi.org/10.1002/14651858.CD001904.pub2

Pellock, J. M., Arzimanoglou, A., D'Cruz, O., Holmes, G. L., Nordli, D., Shinnar, S., & Pediatric Epilepsy Academic Consortium for Extrapolation. (2017). Extrapolating evidence of antiepileptic drug efficacy in adults to children ≥2 years of age with focal seizures: The case for disease similarity. *Epilepsia*, *58*(10), 1686–1696. https://doi.org/10.1111/epi.13859

Pinheiro, M. S., Prado, H. A. do, Ferneda, E., & Ladeira, M. (2015). An Approach for Text Mining Based on Noun Phrases. In *Intelligent Decision Technologies* (pp. 525–535). Springer, Cham. https://doi.org/10.1007/978-3-319-19857-6_45

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Ramaratnam, S., Panebianco, M., & Marson, A. G. (2016). Lamotrigine add-on for drug-resistant partial epilepsy. *The Cochrane Database of Systematic Reviews*, (6), CD001909. https://doi.org/10.1002/14651858.CD001909.pub2

Reiss, W. G., & Oles, K. S. (1996). Acetazolamide in the treatment of seizures. *The Annals of Pharmacotherapy*, *30*(5), 514–519. https://doi.org/10.1177/106002809603000515

Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. *Computational and Mathematical Methods in Medicine*, *2017*, 5140631. https://doi.org/10.1155/2017/5140631

Singh, S. P. (2015). *Quantitative analysis on the origins of morphologically abnormal cells in temporal lobe epilepsy*. University of Cincinnati. Retrieved from https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:ucin1446547280

Singh, S. P. (2016). Advances in Epilepsy: A data science perspective, *58*(2), 89–92. https://doi.org/10.2791/dsj.7.1

Singh, S. P., Chhunchha, B., Fatma, N., Kubo, E., Singh, S. P., & Singh, D. P. (2016). Delivery of a protein transduction domain-mediated Prdx6 protein ameliorates oxidative stress-induced injury in human and mouse neuronal cells. *American Journal of Physiology. Cell Physiology*, *310*(1), C1-16. https://doi.org/10.1152/ajpcell.00229.2015

Singh, S. P., He, X., McNamara, J. O., & Danzer, S. C. (2013). Morphological changes among hippocampal dentate granule cells exposed to early kindling-epileptogenesis. *Hippocampus*, *23*(12), 1309–1320. https://doi.org/10.1002/hipo.22169

Singh, S. P., & Karkare, S. (2017). Stress, Depression and Neuroplasticity. *ArXiv:1711.09536 [q-Bio]*. Retrieved from http://arxiv.org/abs/1711.09536

Singh, S. P., & Karkare, S. (2018, January 5). 10K Pubmed Abstracts related to AntiEpileptic Drugs. https://doi.org/10.6084/m9.figshare.5764524.v1

Singh, S. P., LaSarge, C. L., An, A., McAuliffe, J. J., & Danzer, S. C. (2015). Clonal Analysis of Newborn Hippocampal Dentate Granule Cell Proliferation and Development in Temporal Lobe Epilepsy. *Eneuro*, *2*(6). https://doi.org/10.1523/ENEURO.0087-15.2015

Singh, S. P., Singh, S. P., Fatima, N., Kubo, E., & Singh, D. P. (2008). Peroxiredoxin 6-A novel antioxidant neuroprotective agent. *Neurology*, *70*(11), A480–A481.

Singh, S. P., & Singh, V. P. (2017). Quantitative Analysis on the role of Raffinose Synthase in Hippocampal Neurons. *BioRxiv*. https://doi.org/10.1101/240192

Text Mining Infrastructure in R | Feinerer | Journal of Statistical Software. (n.d.). https://doi.org/10.18637/jss.v025.i05

Trinka, E., Cock, H., Hesdorffer, D., Rossetti, A. O., Scheffer, I. E., Shinnar, S., … Lowenstein, D. H. (2015). A definition and classification of status epilepticus--Report of the ILAE Task Force on Classification of Status Epilepticus. *Epilepsia*, *56*(10), 1515–1523. https://doi.org/10.1111/epi.13121

Turner, A. L., & Perry, M. S. (2017). Outside the box: Medications worth considering when traditional antiepileptic drugs have failed. *Seizure*, *50*, 173–185. https://doi.org/10.1016/j.seizure.2017.06.022

VanStraten, A. F., & Ng, Y.-T. (2012). Update on the management of Lennox-Gastaut syndrome. *Pediatric Neurology*, *47*(3), 153–161. https://doi.org/10.1016/j.pediatrneurol.2012.05.001

Yu, L., Ran, B., Li, M., & Shi, Z. (2013). Gabapentin and pregabalin in the management of postoperative pain after lumbar spinal surgery: a systematic review and meta-analysis. *Spine*, *38*(22), 1947–1952. https://doi.org/10.1097/BRS.0b013e3182a69b90

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(Suppl 13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

Zheng, F., Du, C., & Wang, X. (2015). Levetiracetam for the treatment of status epilepticus. *Expert Review of Neurotherapeutics*, *15*(10), 1113–1121. https://doi.org/10.1586/14737175.2015.1088785

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., … Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, *46*(2), 200–211. https://doi.org/10.1016/j.jbi.2012.10.007
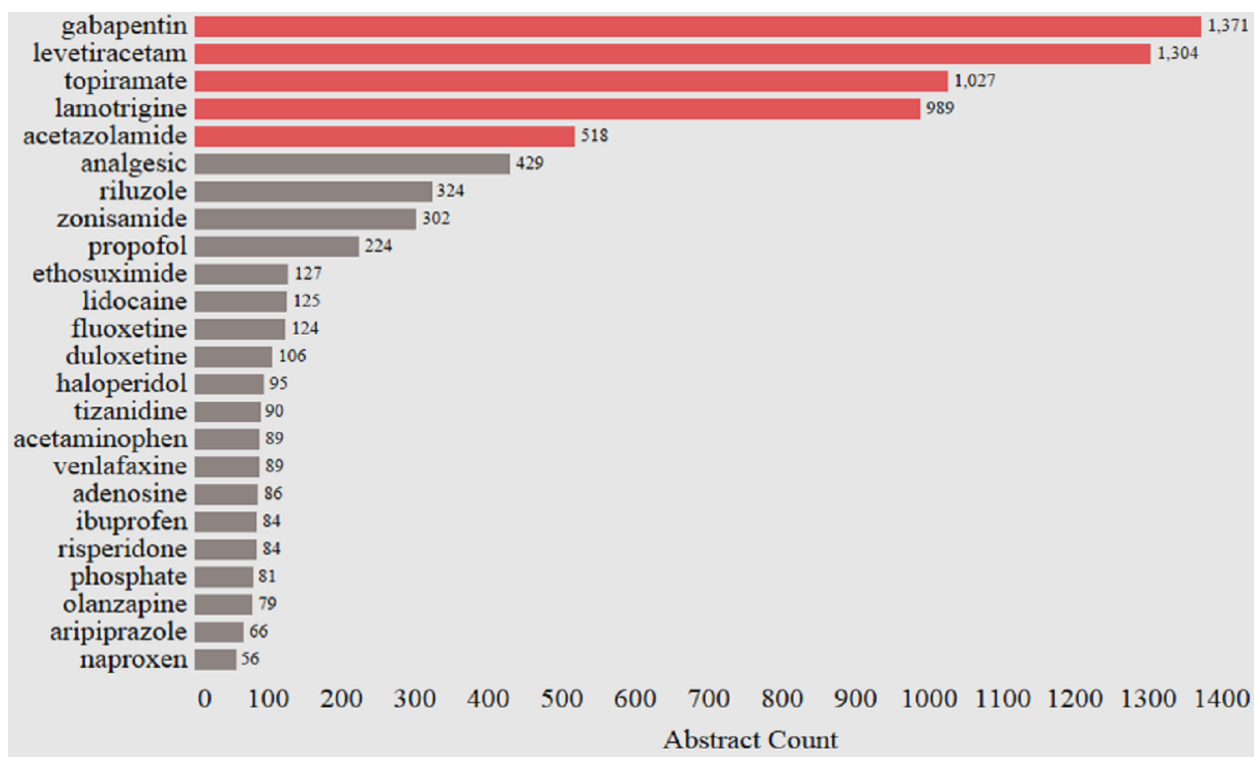
Figure 1: The frequency of PubMed abstracts with anti-epileptic drug mentions from 2007 to 2017. Shown here are the drugs with more than 50 abstracts.
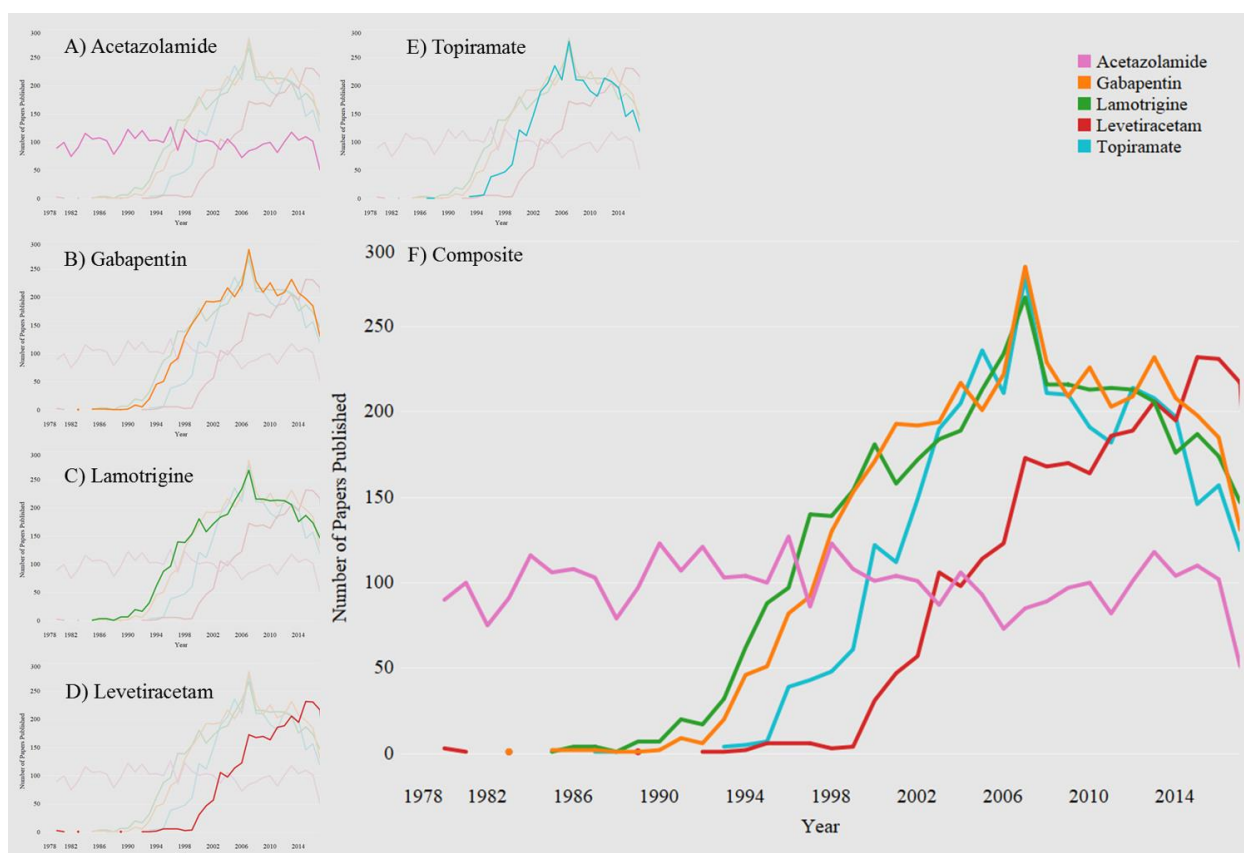
Figure 2: Yearly mentions of the five most frequently occurring anti-epileptic drugs. The frequency of abstracts published between 1980 to 2017 for A) Acetazolamide, B) Gabapentin, C) Lamotrigine, D) Levetiracetam and E) Topiramate and F) Composite of all the drugs (A through E)

Mode of Action	Drug	
Anti-epileptic Agent [EPC],Decreased Central Nervous System Disorganized Electrical Activity [PE]	gabapentin	1,371
	ethosuximide	127
Decreased Central Nervous System Disorganized Electrical Activity [PE]	levetiracetam	1,304
Decreased Central Nervous System Disorganized Electrical Activity [PE],Cytochrome P450 3A4 Inducers [MoA],Cytochrome P450 2C19 Inhibitors [..	topiramate	1,027
Anti-epileptic Agent [EPC],Decreased Central Nervous System Disorganized Electrical Activity [PE],Mood Stabilizer [EPC]	lamotrigine chewable dispersible	989
Carbonic Anhydrase Inhibitor [EPC],Carbonic Anhydrase Inhibitors [MoA],Sulfonamides [Chemical/Ingredient]	acetazolamide	518
Benzothiazole [EPC],Benzothiazoles [Chemical/Ingredient]	riluzole	324
Anti-epileptic Agent [EPC],Decreased Central Nervous System Disorganized Electrical Activity [PE],Sulfonamides [Chemical/Ingredient],Carbonic An..	zonisamide	302
General Anesthesia [PE],General Anesthetic [EPC]	propofol	234
	etomidate	28
Atypical Antipsychotic [EPC]	risperidone	84
	olanzapine	79
	aripiprazole	66
Cyclooxygenase Inhibitors [MoA],Nonsteroidal Anti-inflammatory Compounds [Chemical/Ingredient],Nonsteroidal Anti-inflammatory Drug [EPC]	ibuprofen	84
	naproxen	56
	celecoxib	43
	indomethacin	37
Norepinephrine Uptake Inhibitors [MoA],Serotonin and Norepinephrine Reuptake Inhibitor [EPC],Serotonin Uptake Inhibitors [MoA]	duloxetine	106
	venlafaxine hydrochloride	89
Serotonin Reuptake Inhibitor [EPC],Serotonin Uptake Inhibitors [MoA]	fluoxetine hydrochloride	124
	escitalopram	21
Amide Local Anesthetic [EPC],Amides [Chemical/Ingredient],Antiarrhythmic [EPC],Local Anesthesia [PE]	lidocaine	125
Typical Antipsychotic [EPC]	haloperidol lactate	95
Adrenergic alpha2-Agonists [MoA],Central alpha-2 Adrenergic Agonist [EPC]	tizanidine	90
Adenosine Receptor Agonist [EPC],Adenosine Receptor Agonists [MoA]	adenosine	86
Alkylating Activity [MoA],Alkylating Drug [EPC]	temozolomide	37
Increased Diuresis at Loop of Henle [PE],Loop Diuretic [EPC]	furosemide	36
Serotonin 1b Receptor Agonists [MoA],Serotonin 1d Receptor Agonist [MoA],Serotonin-1b and Serotonin-1d Receptor Agonist [EPC]	sumatriptan	33
Androgen [EPC],Androgen Receptor Agonists [MoA],Androstanes [Chemical/Ingredient]	testosterone	32
Aminoketone [EPC],Dopamine Uptake Inhibitors [MoA],Increased Dopamine Activity [PE],Increased Norepinephrine Activity [PE],Norepinephrine U..	bupropion hydrochloride	28
Calcium Channel Antagonists [MoA],Dihydropyridine Calcium Channel Blocker [EPC],Dihydropyridines [Chemical/Ingredient]	nifedipine	28
Central Nervous System Stimulation [PE],Increased Sympathetic Activity [PE],Sympathomimetic-like Agent [EPC]	modafinil	27
Benzodiazepine [EPC],Benzodiazepines [Chemical/Ingredient]	temazepam	26
Cephalosporin Antibacterial [EPC],Cephalosporins [Chemical/Ingredient]	ceftriaxone	25
Peroxisome Proliferator Receptor alpha Agonist [EPC],Peroxisome Proliferator Receptor gamma Agonist [EPC],Peroxisome Proliferator-activated Rec..	pioglitazone	22
Serotonin 3 Receptor Antagonists [MoA],Serotonin-3 Receptor Antagonist [EPC]	ondansetron hydrochloride	21

Figure 3 The most common drug classes found in the abstract with the frequency of their occurrence. Drug classification was obtained from the US FDA classification.
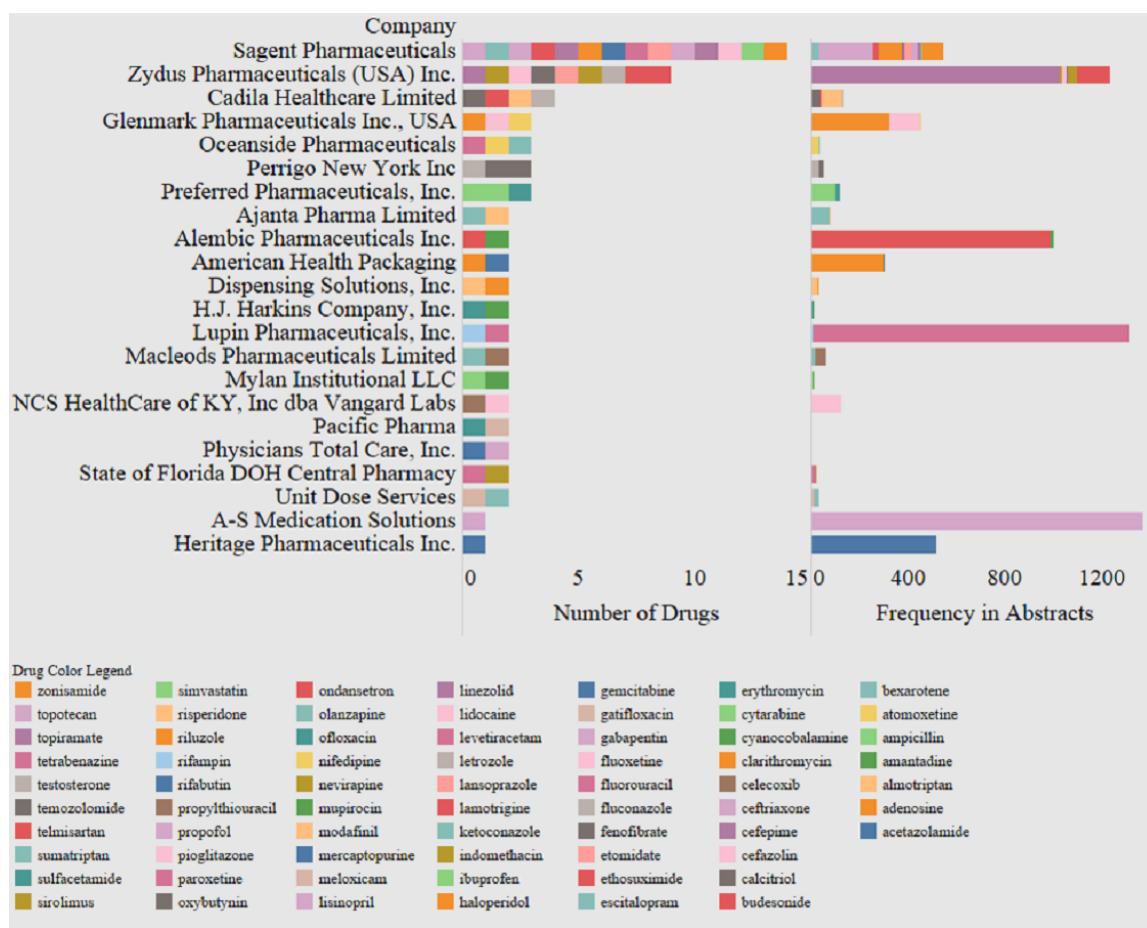
Figure 4. Pharmaceutical companies which have the most number of drugs mentioned in the last 10000 abstracts. The registered name of the company is shown in the first pane, the number of drugs that made it into the study criterion in the second pane and the frequency of drug appearance in the abstracts in the third pane.

Table 1. Sentiment analysis on abstracts with AED mentions. The sentiment of all the abstracts (not-normalized) is shown as either positive (score>0) or negative (score<0). The normalized sentiment scores (score-mean/S.D.) is shown in the right most column. The mean sentiment score was 0.01320 and the standard deviation was 0.2772.

Drug	Sentiment Polarity	Sentiment	Normalized Sentiment
Clopidogrel	Positive	0.002897016	-0.037168052
Mirtazapine	Positive	0.005795628	-0.026711299
Topotecan	Positive	0.009118557	-0.01472382
Fenofibrate	Positive	0.01211505	-0.003913961
Methazolamide	Positive	0.01266136	-0.001943146
Gemcitabine	Positive	0.01526285	0.007441739
Cimetidine	Positive	0.01862922	0.019585931
Cilostazol	Positive	0.02063282	0.026813925
Leflunomide	Positive	0.02083333	0.027537266
Fluoxetine	Positive	0.02222189	0.032546501
Epinephrine	Positive	0.02343064	0.036907071
Loratadine	Positive	0.02425356	0.039875758
Sulfacetamide	Positive	0.02611344	0.046585281
Antibacterial	Positive	0.03498982	0.078606854
Fluorouracil	Positive	0.039489	0.094837662
Ribavirin	Positive	0.03993905	0.096461219
Testosterone	Positive	0.04173582	0.102943074
Gatifloxacin	Positive	0.04434233	0.112346068
Decitabine	Positive	0.04531584	0.115858009
Rifabutin	Positive	0.04868627	0.128016847
Tetrabenazine	Positive	0.05354119	0.145530988
Pioglitazone	Positive	0.05555496	0.152795671
Carisoprodol	Positive	0.05842708	0.163156854
Cytarabine	Positive	0.05848246	0.163356638
Misoprostol	Positive	0.0597341	0.167871934
Omeprazole	Positive	0.06585542	0.189954618
Budesonide	Positive	0.06689463	0.193703571
Calcitriol	Positive	0.06788645	0.197281566
Ketoconazole	Positive	0.08378791	0.25464614
Venlafaxine	Positive	0.0912101	0.281421717
Ganciclovir	Positive	0.09387124	0.291021789
Clarithromycin	Positive	0.09424993	0.292387915
Sumatriptan	Positive	0.1025025	0.322159091
Rifampin	Positive	0.1035337	0.325879149
Piroxicam	Positive	0.1071953	0.339088384
Temozolomide	Positive	0.1075685	0.340434704
Almotriptan	Positive	0.1216465	0.39122114

Itraconazole	Positive	0.1348309	0.438783911
Linezolid	Positive	0.1455073	0.477299062
Sirolimus	Positive	0.148724	0.488903319
Voriconazole	Positive	0.1552595	0.512480159
Paroxetine	Positive	0.1920355	0.645149711
Oxybutynin	Positive	0.20757	0.701190476
Escitalopram	Positive	0.2461041	0.840202381
Propylthiouracil	Positive	0.3178791	1.099130952
Oxandrolone	Positive	0.3825442	1.332410534
Mercaptopurine	Positive	0.4220309	1.474858947
Chlorzoxazone	Positive	0.5641749	1.987643939
Telmisartan	Positive	1.026542	3.655634921
Lisinopril	Positive	2.534934	9.097164502
Meloxicam	Negative	-0.6005493	-2.214102814
Ethosuximide	Negative	-0.2520805	-0.957000361
Hemorrhoidal	Negative	-0.2358386	-0.898407648
Lansoprazole	Negative	-0.2164586	-0.828494228
Cefazolin	Negative	-0.2010915	-0.773057359
Laxative	Negative	-0.2001804	-0.769770563
Expectorant	Negative	-0.1859978	-0.718606782
Menthol	Negative	-0.1659774	-0.646383117
Cefepime	Negative	-0.1627273	-0.634658369
Celecoxib	Negative	-0.1516571	-0.594722583
Olanzapine	Negative	-0.1500488	-0.588920635
Bacitracin	Negative	-0.1498865	-0.588335137
Atomoxetine	Negative	-0.1445815	-0.56919733
Bexarotene	Negative	-0.1382776	-0.546455988
Bupropion	Negative	-0.1250089	-0.498589105
Lidocaine	Negative	-0.1211533	-0.484680014
Cyanocobalamine	Negative	-0.1130527	-0.455457071
Ampicillin	Negative	-0.1123144	-0.452793651
Indomethacin	Negative	-0.1115034	-0.449867965
Carboplatin	Negative	-0.1075859	-0.43573557
Gabapentin	Negative	-0.107477	-0.435342713
Amantadine	Negative	-0.105775	-0.429202742
Oxaliplatin	Negative	-0.1030068	-0.41921645
Aspirin	Negative	-0.1023196	-0.416737374
Ceftriaxone	Negative	-0.1000455	-0.40853355
Diphenhydramine	Negative	-0.09806558	-0.401390981
Cortisone	Negative	-0.09325419	-0.384033874
Modafinil	Negative	-0.09214975	-0.380049603
Doxycycline	Negative	-0.08559167	-0.356391306

Clotrimazole	Negative	-0.0839393	-0.350430375
Metronidazole	Negative	-0.08369748	-0.349558009
Letrozole	Negative	-0.08360991	-0.3492421
Paclitaxel	Negative	-0.08267615	-0.345873557
Topiramate	Negative	-0.08196571	-0.343310642
Lamotrigine	Negative	-0.07665585	-0.324155303
Temazepam	Negative	-0.07443998	-0.316161544
Azithromycin	Negative	-0.07425097	-0.31547969
Hydrocortisone	Negative	-0.07420852	-0.315326551
Valsartan	Negative	-0.0687518	-0.295641414
Ciprofloxacin	Negative	-0.06780962	-0.292242496
Adenosine	Negative	-0.06476878	-0.281272655
Risperidone	Negative	-0.06339783	-0.276326948
Etomidate	Negative	-0.05926828	-0.261429582
Aripiprazole	Negative	-0.05923076	-0.261294228
Nicotine	Negative	-0.05807075	-0.257109488
Zaleplon	Negative	-0.0561836	-0.250301587
Duloxetine	Negative	-0.05339635	-0.240246573
Tizanidine	Negative	-0.05224761	-0.236102489
Acetaminophen	Negative	-0.05086406	-0.231111328
Ropinirole	Negative	-0.0489206	-0.224100289
Hydrochlorothiazide	Negative	-0.04238263	-0.200514538
Levetiracetam	Negative	-0.04194175	-0.198924062
Phosphate	Negative	-0.04153744	-0.197465512
Nevirapine	Negative	-0.04135263	-0.19679881
Guanfacine	Negative	-0.03962207	-0.190555808
Isoniazid	Negative	-0.0354021	-0.175332251
Simvastatin	Negative	-0.03430078	-0.171359235
Acetazolamide	Negative	-0.0340079	-0.17030267
Propofol	Negative	-0.03290462	-0.166322583
Ondansetron	Negative	-0.03218649	-0.163731926
Ofloxacin	Negative	-0.03167064	-0.161870996
Levofloxacin	Negative	-0.02956457	-0.154273341
Zonisamide	Negative	-0.02831881	-0.149779257
Riluzole	Negative	-0.02446827	-0.13588842
Nifedipine	Negative	-0.02092019	-0.123088709
Furosemide	Negative	-0.01862298	-0.114801515
Ibuprofen	Negative	-0.01638219	-0.106717857
Naproxen	Negative	-0.01059067	-0.085824928
Fluconazole	Negative	-0.007489149	-0.07463618
Eszopiclone	Negative	-0.006686009	-0.071738849
Haloperidol	Negative	-0.005274025	-0.066645112

Disposable	Negative	-0.004070103	-0.062301959
Minoxidil	Negative	-0.001773493	-0.05401693
Lovastatin	Negative	-0.001518158	-0.053095808
Acyclovir	Negative	-0.001015911	-0.05128395
Erythromycin	Negative	-0.00086424	-0.050736798

Table 2. Results from mLDA based Topic models run on abstract containing the top 5 drugs. Each topic is represented by the top keywords defining the topic.

Gabapentin		Levetiracetam		Topiramate		Acetazolamide	Lamotrigine
Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 1
Pain	Analog	Refractory	Indicated	LGS	Controlling	Clinical	Blood
Group	Incidence	Generalized	Woman	Epilepsy	Drug Resistant	Macular	Dried
Controlled	Scores	Seizures	Partial Epilepsy	Lennoxgastaut	Effective	Visual	Spots
Spinal	Inclusion	New	Response	Drugs	Long Term	Acuity	Concentrations
Surgery	Scale	Brivaracetam	Therapy	Resolution	Drop	Better	Accuracy