CSE 6242: Final Report
Team #16: Medicines in the Media

Bahl, Vikram
Baumann, Christopher
Goodyear, Trevor
Hilton, Charity

Dec 8 2017

1 Introduction and Motivation

Researchers are increasingly utilizing large healthcare databases to study the benefits and harms of medications for patients with a variety of conditions. One challenge of using these data sets is patients are often exposed to information outside of the medical setting that may affect their decisions regarding a treatment. For example, patients may stop taking their cholesterol medications after reading about the dangers of statin drugs, or request a prescription for Tamiflu because of fears over the bird flu epidemic. However, little has been done to examine this phenomenon.

In the current study, we examined these clinical and non-clinical influences on prescription drug behavior in the United States. Currently, there is no work linking media coverage on specific treatments to how they are being prescribed to, and consumed by, patients. This connection provides value to decision-makers who wish to be more informed, and to medical professionals interested in clinical decision support to provide the best care for their patients. Pharmaceutical companies are also key stakeholders as they in the development of medications and treatments, but may be wary of using large data sets for assessing medication effects. By using publicly available data sets on pharmaceuticals in conjunction with coverage in both news and social media, we created an interactive analysis tool to link news media mentions with medication prescription trends in the United States between 1992 and 2017.

2 Problem Definition

Prescription drug and information about diagnoses are influenced by popular and social media sources. However, there are few resources to evaluate and measure their correlation.

3 Literature Survey

Medical decision making is strongly influenced by media coverage. However, reports often contain misinformation or a lack of information[6][8][11], which has been directly associated with negative outcomes for patients[10]. This is especially important as over half of adults receive healthcare information from online sources[4]. Direct-to-consumer advertising has also led to a rise in prescription requests and fulfillment of advertised products[7][12]. Similarly, coverage of high-profile cases of illnesses and treatments has had a measurable effect on the public. For example, news coverage of the death of Katie Courics husband from colon cancer brought about an increase in colonoscopy rates in the months following[2]. The media can even affect medical decision making without making claims or recommendations. For example, a health alert in Canada on the potential adverse effects of antidepressant use in pregnant mothers on newborns influenced the mothers choice to continue taking the medication, despite no specific recommendations in the alert[3]. Overall, the media has been shown to have a strong influence on individual medical decision making, but little has been done to examine this phenomenon on a large scale for pharmaceuticals; a gap we intend to investigate.

Based on the influence of media on medical decision making, researchers in pharmacovigilance, the study of adverse effects of pharmaceutical products, have previously looked at social media to detect adverse drug reactions (ADRs) using various forms of textual analysis[1][9]. One study used a Bidirectional Long Short-Memory LSTM (BLSTM) Recurrent Neural Network (RNN) to recognize tweets on Twitter as a series of tokens which are used to predict whether an ADR occurred. The tokens were mapped to the pre-trained word2vec model[5] to predict the label for each token.

4 Proposed Method

4.1 Intuition

Our project is a new approach because it looks across all medications from Medicaid data from 1992 to 2017. It links popular and social media mentions over time, using sources such as Twitter, New York Times, and the Guardian,

using text mining techniques such as sentiment analysis. Some of these techniques have been applied on medical data but mostly on a case-by-case basis.

4.2 Description

The following sections detail our approaches:

4.2.1 Data Collection and Cleaning

This project used existing data sets from the Centers for Medicare & Medicaid Services (CMS) on the medications prescribed by health care providers covered by Medicaid in order to analyze the prescription trends over the last 25 years (since 1992). This data was cleaned to map medications to standardized vocabularies (RxNorm) and normalized by states and years. Recent Medicaid data sets have on average 4 million rows per year, while the latest Medicare data set has 24.5 million rows per year.

Our data set for drug names came from the RxNorm dictionary, which is an openly available data set from the National Library of Medicine. We used the OMOP common data model to extract things like brand names and drug classes.

We also scraped The New York Times and The Guardian using their APIs for mentions of particular pharmaceuticals that appear within the CMS data sets. In addition, we scraped Twitter for mentions of the drugs listed for possible ADRs.

To provide another dimension of data and quantify adverse events, we scraped the FDA Adverse Event Reporting System (FAERS). This data is available from 2014 to 2017 and provides a quantified metric of serious reactions of medications reported to the FDA.

The following is a summary of our data sources:

1. Centers for Medicare & Medicaid Services (CMS) Medicaid State Drug Utilization data sets for 1992-2017 (74M records, temporal)
2. National Library of Medicine RxNorm Dictionary, OMOP v5 Vocabulary (37.8k event tables)
3. FDA Adverse Event Reporting System data set (2.7M unique events, temporal)
4. Custom query using The New York Times Article Search API and
5. The Guardian Open Platform API (28.4k articles, temporal)
6. Twitter REST API (229k tweets, temporal)

4.2.2 System Architecture

System Components The final application technology stack consists of the following persistent services. MongoDB was used as an intermediate store for collected tweets and Adverse Drug Reaction results. The results from MongoDB were moved to PostgreSQL in the final data setup.

1. Static web application files (HTML, Javascript, CSS)
2. Python Flask API
3. PostgreSQL

The Python Flask API is served through Apache with `mod_wsgi`. Apache also serves as a transparent proxy for the static web application files, which are served through NodeJS. NodeJS serves the web files in order to facilitate automatic recompilation following frontend code changes.

Azure Environment The first attempt at a system architecture utilized Azure cloud services. These services included an Ubuntu virtual machine for the web application and API, an Azure SQL instance, and an Azure CosmosDB instance. Issues were encountered with both the Azure SQL and CosmosDB instances.

The SQL database contained approximately 75 million rows. The pricing model for Azure SQL includes limiting the amount of "database transactions per second." This limit is catastrophic for any SQL operations which require a full table scan. This could have been mitigated by the explicit creation of lookup tables for every query type that this application requires.

Azure CosmosDB is the Azure implementation of a NoSQL document database. This product mostly conforms to the MongoDB API but it lacks some key features. Specifically, full-text indexes are not supported. Azure recommends the use of the Azure Search service which is not real time and requires an extra subscription that can be as expensive as $1250/mo.

Since Azure could not be utilized effectively for this project, a DC/OS cluster in some of the authors' research lab was utilized for the final application. DC/OS is a container orchestration platform and proved to be an adequate host for the PostgreSQL instance.

4.2.3 Data Summarization and APIs

The following API endpoints are provided through Flask. All endpoints except for `/sentiment/` accept a drug name and return the associated data set for that drug.

`GET /tweets/<drugname>` This endpoint returns tweets that contain the given drug name. The corpus of tweets that this endpoint searches is a Twitter Stream executed for approximately two weeks on the top 125 most popular drugs, as measured by overall prescription volume.

`GET /counts/<drugname>/` This endpoint returns a per-year prescription count for all years in which the drug was prescribed, as indicated by the Medicaid data.

`GET /countsBreakdown/<drugname>/` This endpoint returns a per-state/year/quarter prescription count for all years in which the drug was prescribed, as indicated by the Medicaid data.

`GET /adr/<drugname>/` This endpoint returns Adverse Drug Reactions as detected through the process described in Section 4.2.4.

`GET /events/<drugname>/` This endpoint returns all FDA Approvals and scraped News Articles for the given drug.

`GET /faers/<drugname>/` This endpoint returns all FAERS outcome events for the given drug.

`POST /sentiment/` The `/sentiment/` endpoint accepts a JSON list of strings and for each string it computes a sentiment score using NLTK's VADER library.

4.2.4 Twitter Event Detection

For detecting Adverse Drug Reaction (ADR) from the Tweets, we replicated the state of the art deep learning capability of sequence labelling the words in a Tweet as a ADR or not. This was performed by recreating Bidirectional Long Short-Term Memory LSTM (BLSTM) Recurrent Neural Network (RNN) to detect ADRs on Twitter and expand coverage to include new drugs to create a community driven dictionary of reactions[1].

4.2.5 Sentiment Analysis

We applied sentiment analysis to scrap text data from the New York Times, The Guardian and Twitter. For this analysis, we used the Python NLTK library, Vader. This library assigns sentiment scores to text in positive, negative and neutral categories. We displayed these categories in the histogram to provide another data dimension.

4.2.6 User Interface

The user interface is built as a web application, primarily using the ReactJS library. ReactJS is useful for creating componentized modern user interfaces. Each component is responsible for its own state, and thus the view and its model remain coupled, so that the component can easily be moved and refactored. React manages components in a Virtual DOM, before any changes are made to the actual DOM, thus React is know to be very performant. In addition, it uses Bootstrap CSS in 'Lux' style. The application uses Babel to compile JSX to JavaScript, webpack to bundle dependencies and Node to manage dependencies and as a web server. The application has 4 major components:

- **Autocomplete:** Interactive autocompleter to filter and select medication name.
- **Medication Detail:** Panel to display medication images, descriptions, adverse event mentions filtered as a treemap, media coverage, and twitter mentions
- **Map:** Heat map of the United States relative counts for prescriptions, filtered by medication and year.
- **Timeline:** Interactive timeline to filter date events with major events displayed such as FDA approval date. Playback of all time events, updating maps and medication summary.
- **Medication Summary:** Interactive histograms to display medication counts, media coverage articles and adverse event by time.

The main goal of the user interface was to perform interactive longitudinal prescription data exploration linking social and news media.

5 Experiments / Evaluation

5.1 Description

To evaluate our application, we conducted a survey to 20 participants to evaluate if interacting with the application had a tangible effect on their perception on 10 common medications, including some available over-the-counter as well as ones only by prescription.

5.2 Details

The survey respondents were asked the following question, with a likert scale to make responses.

1. How likely is the information you now have going to impact your decisions regarding that drug?
   (a) Very Unlikely
   (b) Somewhat Unlikely
   (c) Neutral
   (d) Somewhat Likely
   (e) Very Likely

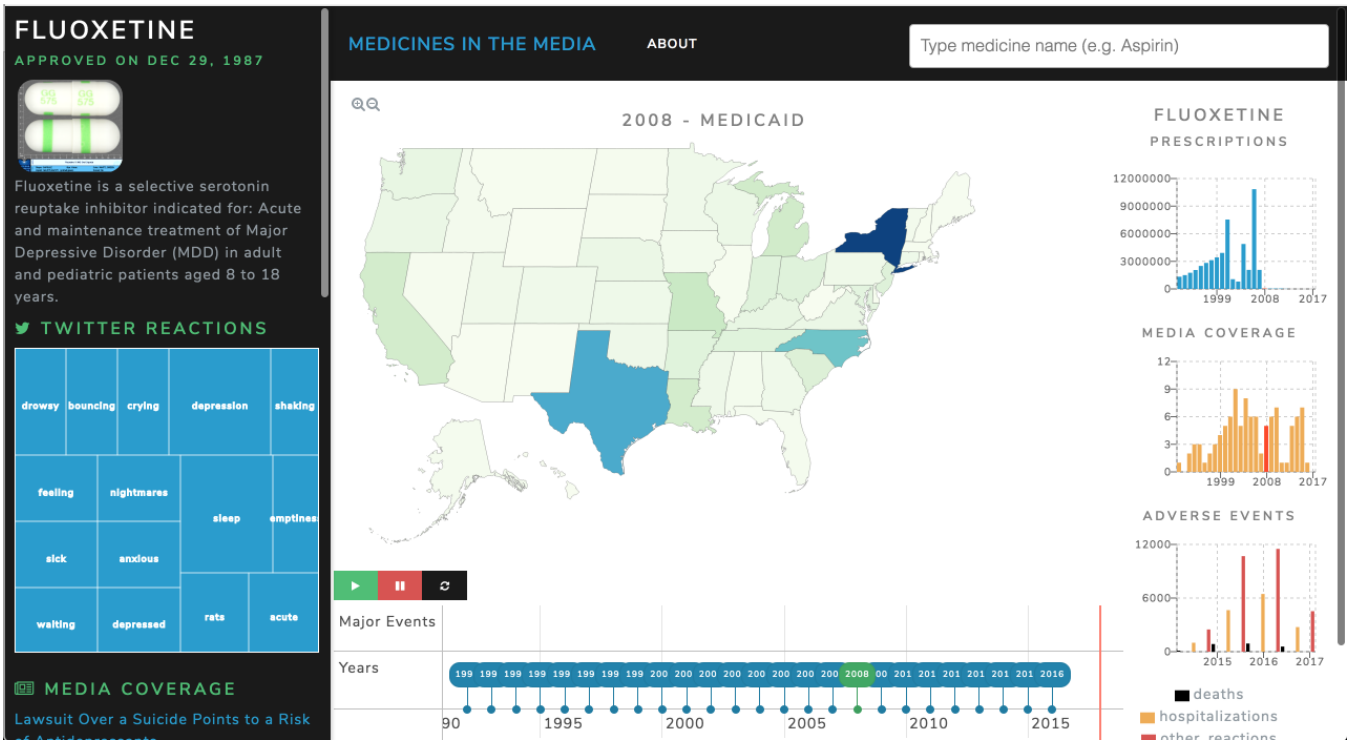

Figure 1: Screen capture of Medicines in Media with Fluoxetine (Prozac) selected

(Note: After this question, users were given the following prompt, "Now interact with our application. Use this list of drugs as a reference while browsing through the information. Continue to the next question when you're ready. Please note, this will open in a new tab so you may need to disable your pop-up blocker.")

2. After viewing the information presented in the application, how has your opinion of the drug changed?
   (a) Less Favorable
   (b) Somewhat Less Favorable
   (c) Neutral
   (d) Somewhat More Favorable
   (e) More Favorable
3. How likely is the information you now have going to impact your decisions regarding that drug?
   (a) Very Unlikely
   (b) Somewhat Unlikely
   (c) Neutral
   (d) Somewhat Likely
   (e) Very Likely
4. What is your age group?
5. What is your gender?
6. What is your household income?

7. What ethnicity do you identify as?

The users were asked to evaluate on the following drugs (with additional detail provided).

- Acetaminophen
- Hydrochlorothiazide
- Amitriptyline
- Glyburide
- Atorvastatin
- Codeine
- Guaifenesin
- Amoxicillin
- Pseudoephedrine
- Fluoxetine

5.3 Observations

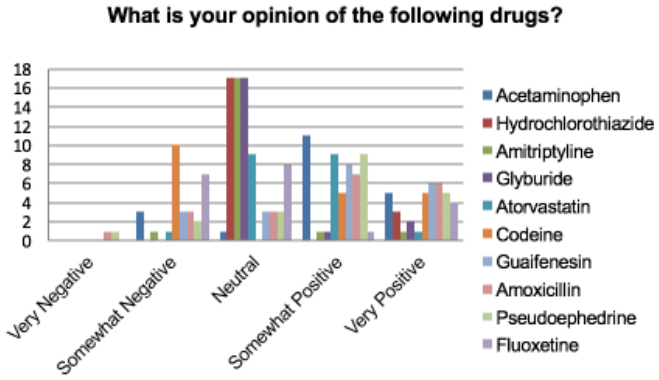

Figure 2: Results before interacting with Medicines in the Media

Depending on the drug, the evaluation showed a positive or negative impact on how favorably people viewed drugs. Our findings show that there is a correlation between prescribing trends and media coverage as well as a negative correlation with confirmed adverse events. The results from our evaluation indicate that while 16% of participant responses were negative, 25% stated their opinion decreased after viewing the presented information. The results further show that publicly accessible information can have an effect on their opinion of a prescription drug, as 54% of participants indicated a change in opinion.

The survey is currently hosted at https://gatech.co1.qualtrics.com/jfe/form/SV_0AkU29G1Fk0uN6t.

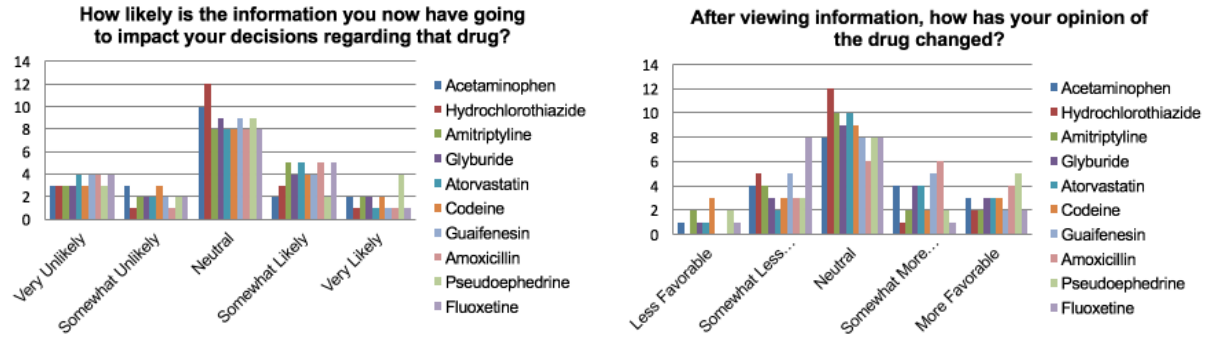

Figure 3: Results after interacting with Medicines in the Media

6 Conclusion

6.1 Summary

Patients have access to more information than ever before when making decisions regarding their health. Despite our initial assumption that this tool would be useful for health researchers, we believe this is a valuable tool for consumers as over half of the evaluation participants expressed a change in their attitude towards the selected drugs after interacting with our application. While some trends in prescription drug behavior were identified to be correlated with media mentions and adverse events, there are opportunities for expanding the work done in the project. Further data inputs are needed to test the power of media mentions, things like cost, policy, and new drugs coming on the market. Inconsistent Medicaid data reporting may have also skewed results. Exploring these areas would provide a greater explanation for the impacts on the trends in prescribing behavior.

6.2 Opportunities for Expansion

6.2.1 Data

The primary opportunity for expansion is new data sources for prescription data. Medicaid data is useful in that it provides a larger span of years (1992-2017) versus Medicare (2010-2012). However, Medicaid data lacks granularity of smaller geographic areas such as county or city. In addition, most government data sets may be biased to certain populations (such as older patients). A comprehensive survey of prescription counts from public and private insurers would provide better results.

In addition, it would be valuable to scrape additional media sources such as television broadcast media and a longer period of social media data. It may also prove valuable to include other variables, such as legislation, weather, and stock market values.

6.2.2 Time Series Correlations

In addition, as additional events are discovered and entered, the next level of evaluation would be time series correlation to determine if certain events can be mathematically linked to increases and decreases to number of medications prescribed.

6.3 Application Demonstration

The application is currently deployed at medicine.usnewsmap.com. Individual medications can also be queried at medicine.usnewsmap.com/?q=fluoxetine.

6.4 Distribution of Team Effort

Task	Member(s) Responsible
Back-end: Medical Data Collection & Cleaning	Charity & Trevor
Back-end: Twitter Data Collection & Cleaning	Trevor & Vikram
Back-end: Devops and Server Setup	Trevor
Back-end: Develop Database and Sentiment Analysis APIs	Trevor & Vikram
Back-end: Develop and Deploy Event Categorization Method	Vikram
Back-end: News Media Scraping	Chris
Front-end: UI Design	Charity
Front-end: UI Development: Events, Visualizations and API hooks	Charity
Evaluation: Create Evaluation Survey	Chris
Quality Assurance, Evaluation, & Refinements	All
Milestone: Project Proposal	All
Milestone: Project Proposal Presentation	Charity
Milestone: Create Progress Report	Chris
Milestone: Create Project Poster	Chris
Milestone: Poster Presentation	All
Milestone: Create Final Report	All

References

[1] A Cocos, AG Fiks, and AJ Masino. "Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts". In: *Journal of the American Medical Informatics Association* 24.4 (2017), pp. 813–821. DOI: 10.1093/jamia/ocw180.

[2] P Cram et al. "The impact of a celebrity promotional campaign on the use of colon cancer screening: The Katie Couric effect". In: *Archives of Internal Medicine* 163.13 (2003), pp. 1601–1605. DOI: 10.1001/archinte.163.13.1601. URL: http://dx.doi.org/10.1001/archinte.163.13.1601.

[3] A Einarson et al. "SSRI'S and other antidepressant use during pregnancy and potential neonatal adverse effects: Impact of a public health advisory and subsequent reports in the news media". In: *BMC Pregnancy and Childbirth* 5.1 (2005). DOI: 10.1186/1471-2393-5-11.

[4] S Fox. "The Social Life of Health Information". In: *Pew Internet & American Life Project* (May 2011). URL: http://www.pewinternet.org/files/old-media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf.

[5] F Godin et al. "Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations". In: *Proceedings of the Workshop on Noisy User-generated Text* (2015). DOI: 10.18653/v1/w15-4322.

[6] A Larsson et al. "Medical messages in the media - barriers and solutions to improving medical journalism". In: *Health Expectations* 6.4 (2003), pp. 323–331. DOI: 10.1046/j.1369-7625.2003.00228.x.

[7] B Mintzes et al. "How does direct-to-consumer advertising (DTCA) affect prescribing? A survey in primary care environments with and without legal DTCA". In: *Canadian Medical Association Journal* 169.5 (2003), pp. 405–412. eprint: http://www.cmaj.ca/content/169/5/405.full.pdf+html. URL: http://www.cmaj.ca/content/169/5/405.abstract.

[8] R Moynihan et al. "Coverage by the News Media of the Benefits and Risks of Medications". In: *New England Journal of Medicine* 342.22 (2000), pp. 1645–1650. DOI: 10.1056/nejm200006013422206.

[9] A Nikfarjam et al. "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features". In: *Journal of the American Medical Informatics Association* (Sept. 2015). DOI: 10.1093/jamia/ocu041.

[10] PJ Sambrook, BE Nordin, and AN Goss. "Impact of adverse news media on prescriptions for osteoporosis: effect on fractures and mortality". In: *The Medical Journal of Australia* 194.1 (2011), pp. 51–52.

[11] G Schwitzer. "Addressing tensions when popular media and evidence-based care collide". In: *BMC Medical Informatics and Decision Making* 13.Suppl 3 (2013). DOI: 10.1186/1472-6947-13-s3-s3.

[12] MS Wilkes, RA Bell, and RL Kravitz. "Direct-to-consumer prescription drug advertising: trends, impact, and implications". In: *Health Affairs* 19.2 (Jan. 2000), pp. 110–128. DOI: 10.1377/hlthaff.19.2.110.