Real-World Evidence and Social Media: Case Studies

Sreeram Ramagopalan, PhD
Research Scientist, Real-World Evidence, Evidera

Andrew Cox, PhD
Research Scientist, Real-World Evidence, Evidera

Sreeram Ramagopalan

Andrew Cox

The analysis of social media is becoming a powerful tool that is being used increasingly to answer research questions across numerous areas including disease spatiotemporal epidemiology and drug adverse events. Patients are increasingly using web technologies such as social media (e.g., Twitter and Facebook), blogs, and forums to generate and access opinions of diseases and treatments. For rare diseases, patient social media is important to the patient population as it represents one of the few means of contacting others with the condition. There are specific forums for almost every disease and condition. Many contain large volumes of patient posts, posted by a community of hundreds, thousands, or tens of thousands of patients with the discussion records often reaching back several years. Treatments, treatment options, and symptoms are often the largest topic of conversation, with a growing trend for posters to summarize their entire treatment and test history, with dates, in the footer of their postings. Other metadata included in the posts often contains information on join date, posting date and time, and sometimes geolocation. In addition to monitoring adverse drug events, social media can provide a means of mapping the symptoms, treatments, outcomes, and development of rare and less well characterized conditions where published accounts are lacking. This vast volume of content therefore serves as an important potential source of real-world data that can be used for pharmacoepidemiology and other research.

However, there are barriers to effective use of this real-world data, as dealing with text in social media settings is hugely challenging. Different styles of communication, shorthand, typos, spelling mistakes, and contextual meaning all make analytical approaches difficult. For example, within a collection of posts on breast cancer the treatment Docetaxel can be referred to as 'tax', 'doci', 'docy', 'docitaxal', 'dositaxal', 'dositaxel.' Words also often need to be accounted for in context. Using the same example where the 'tax' abbreviation is used, it is necessary to differentiate the statements "yesterday my specialist put me on tax" and "hoping to receive a tax rebate." We present two case studies which attempt to give a flavor of the potential of patient-related social media as a data source.

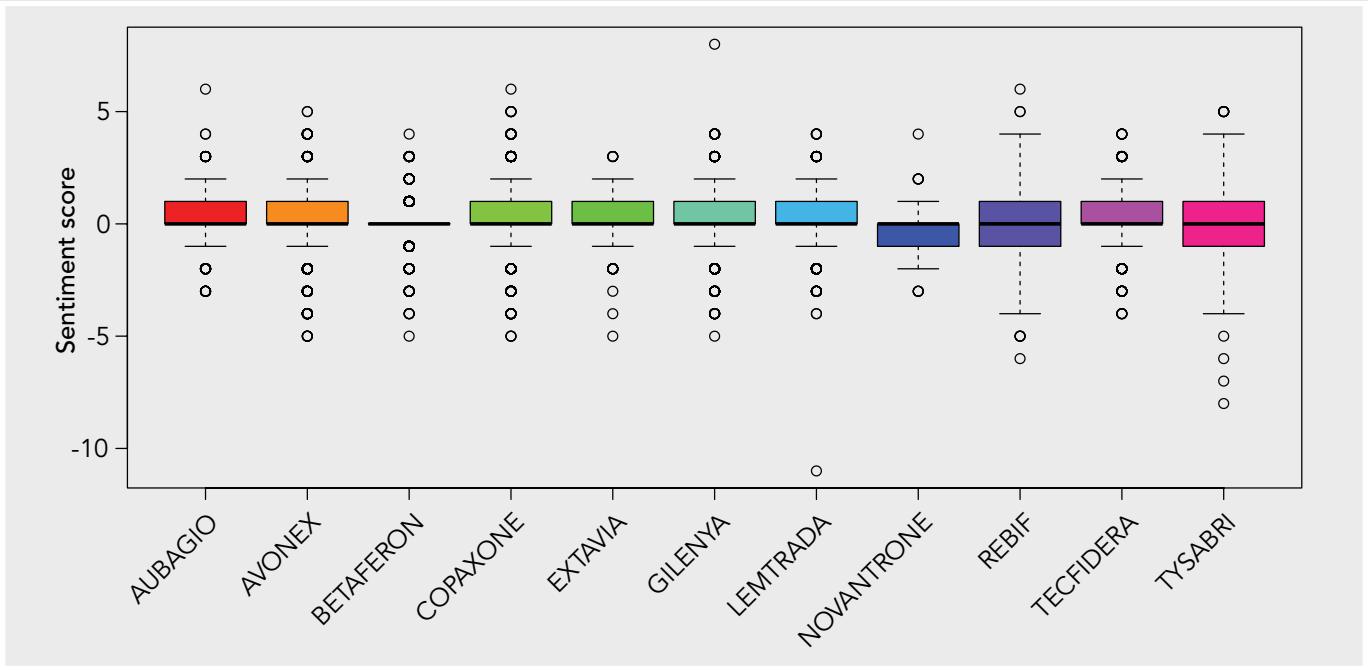
Case Study 1: An Analysis of Tweets about MS

Multiple sclerosis (MS) is a chronic, neurodegenerative autoimmune disorder of the central nervous system and is the most commonly acquired neurological disorder in young adults. Given the age of the patient population and recent availability of many therapies for MS, we explored whether we could analyze social media to help gauge patient opinion about MS treatments.[1]

We used the popular social media site Twitter (http://twitter.com) to explore the reporting of patient opinion about MS treatments. We found approximately 60,000 tweets relating to an MS treatment.

In order to analyze text, *Natural Language Processing* had to be employed. Natural language processing (NLP) is a way for computers to analyze, understand, and derive meaning from human language in a useful way. Because of the short nature of tweets and the presence

Figure 1. Boxplot of sentiment scores of tweets



of typographical errors, ad-hoc abbreviations, phonetic substitutions, ungrammatical structures, and emoticons, NLP was first used to essentially "clean the data" and make better sense of the tweet texts.

NLP was then used to perform sentiment analysis. Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic or product is positive, negative, or neutral. A sentiment score can then be generated - positive scores indicating a preferential statement or negative scores a disapproving one.

In our analysis, we found that about half of all tweets had a neutral sentiment. Combining tweets that contained sentiment showed a significantly different mean sentiment score between drugs (see Figure 1).

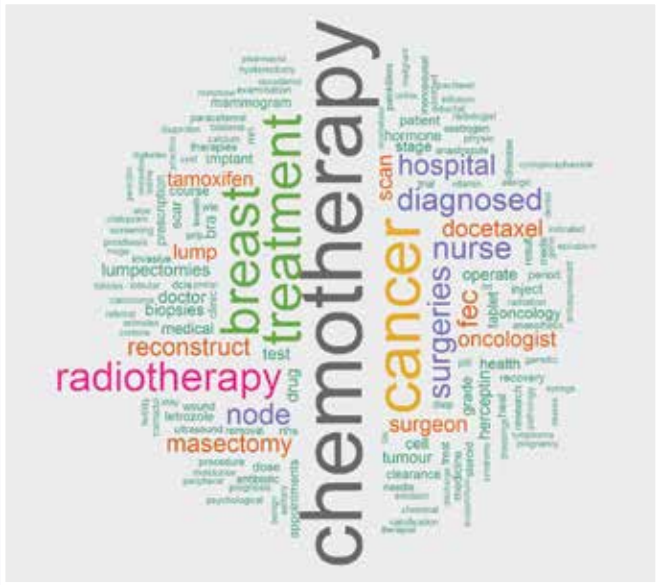
Most common words in tweets for treatments were also investigated and word clouds generated. A word cloud is an image composed of words used in a particular text, in which the size of each word indicates its frequency. An example word cloud for the 50 most common words for one treatment investigated here is shown in Figure 2, highlighting potential adverse events "pml" and disease activity "relapse."

Overall we concluded that a significant proportion of tweets did contain either positive or negative statements about MS treatments, and the distribution of sentiment score was different between treatments. Thus it appears that Twitter can be a potential resource to understand patient opinion about MS treatments. When looking at frequency of words, words known to be associated with particular drugs (e.g., "infusion") were identified providing some face validity for our results reflecting real, specific tweets about MS treatments.

Figure 2. Word cloud for one MS treatment

Figure 3. Word cloud for words tagged "symptom/side effect" related from breast cancer postings

Figure 5. Word cloud for words tagged "medical" related from breast cancer postings



Case Study 2: Characterising Breast Cancer

In this simple analysis, over 120,000 posts spanning several years were collected from leading breast cancer discussion forums. Posts were processed to "tag" emotional words, medical words, and words related to symptoms/side effects. Term heterogeneity (variation in words with the same intended meaning) was accounted for by the creation of word variant dictionaries. Once words were corrected from variants, word frequency tables were produced for each tagged category of word type. Word clouds were produced from the word frequency tables.

Figure 4. Word cloud for words tagged "emotional" related from breast cancer postings

These results represent what can be potentially done relatively quickly and easily using data from Twitter. More rigorous analytical methods can be applied for more specific questions (e.g., the analysis of adverse events, treatment preferences and switches, and investigations of disease epidemiology).

The evolution of healthcare, the high cost of certain drugs, the competition of new drugs entering the market, and the need to show how treatments really work in the real world have all had a part in increasing the importance and value of real-world evidence (RWE). As the demand for RWE grows, the definition and scope of what that evidence entails also grows, and patient input via social media is becoming a very real part of building the RWE story. While the case studies in this article show a fairly simplistic use of social media in this way, the options of its use in representing true value from the patients' perspective are limitless and will continue to evolve. Social media is here to stay, and in many cases is the primary way for patients sharing a common disease, health state, or treatment to share real-world experiences. As such, social media has become a source of real-world evidence that cannot be ignored.

For more information, please contact Sreeram.Ramagopalan@evidera.com or Andrew.Cox@evidera.com.

REFERENCES

[1] Ramagopalan S, Wasiak R, Cox AP. Using Twitter to Investigate Opinions about Multiple Sclerosis Treatments: A Descriptive, Exploratory Study. *F1000Res.* 2014 Sep 10;3:216. doi: 10.12688/f1000research.5263.1. eCollection 2014.