Can Social Media bring more public awareness to rare diseases, leading to more clinical studies?

Geetha Ramaswamy
School of Informatics, Computing, and Engineering
Indiana University Bloomington
USA
gramaswa@iu.edu

ABSTRACT

Social media platforms facilitate the sharing and discussion of topics and have the added advantage of proactively notifying and informing millions of interested users instantaneously as soon as any activity on that topic occurs. Since topics shared in social media are not expected to follow any academic writing standards, anyone is able to share information informally and quickly with the primary intention of relaying the information overriding how the message containing the information is being "worded". As a result, social media disseminates information to every corner of the world, particularly the "common man", bringing about prolific changes impacting the public good, that would otherwise happen at a snail's pace.

A place where social media has made a key difference is in the area of public awareness of rare diseases. This paper studies the impact of public awareness around a rare disease such as IPF on the number of ongoing clinical studies for the disease, quarter to quarter, starting from 2016 until the 3rd quarter of 2017. In doing so, we calculate a social awareness quotient for IPF to quantitatively measure the social awareness. In order to compare this quotient and benchmark it, we also calculate the social awareness quotient of a more publicly well-known disease such as Breast Cancer. From our research, we were able to see a positive correlation between increasing social awareness quotients and the number of clinical studies for that disease.

Given this, by taking the appropriate actions and influencing the social media content, rare disease advocates for most rare diseases can seek to positively influence the increase in clinical studies for the same, thus leading to a better prognosis, orphan drugs and treatment for that disease.

KEYWORDS AND PHRASES

Rare Diseases on Social Media, IPF Awareness, Social Media, IPF, Social Awareness Quotient, Rare Diseases, Disease Awareness, Clinical Studies

1 INTRODUCTION

Ever since the advent of the internet age two decades ago, accessibility to knowledge has only been skyrocketing, benefiting both the general public as well as subject matter experts in almost every industry. Especially with the emergence and adoption of social media, particularly in the last 5 to 7 years, this trend has only been empowered much more. An industry that has specifically benefited from this positive impact would be the healthcare industry, according to Phil Baumann (2009). Given this, in this paper, we will research how much social media has contributed to the awareness around diseases, in particular, ones that are rare in prevalence, and therefore suffer from a poor prognosis due to lack of sufficient studies and research both of which are vital to the determination of treatment and cure for the disease. Taking a rare disease as an example, we will assess whether an improved awareness through social media has contributed to more clinical studies for that disease over the last couple of years.

While there is the internet and Google searches providing us with articles from disparate multiple sources and health organizations, social media, and the Twitter platform in particular, allow these multiple sources to come together as a group, tagging discussions and topics using a common word or "hashtag". Anyone interested in the topic could use Twitter to access the topic's most prominent "hashtags" and follow the groups who contribute to the topic discussion. In other words, the information discussed on this topic is now shared publicly amongst the groups themselves such as healthcare institutions and providers as well as with those impacted directly by those diseases such as the patients and their loved ones. The biggest advantage from this form of information broadcasting in Twitter is its ability to promote an exponentially high level of awareness across the board as information is dispensed without discrimination. In the absence of social media, patients and their loved ones who eagerly pursue information related to disease symptoms, coping techniques, prognosis and treatment options are severely limited to their primary care provider and have no perspective on how any of their co-sufferers outside of their limited circle of family and friends are coping. As the Twitter platform does not require mutual authorization between users, it is also possible, for any patient (or user in Twitter) to send or "tweet" a question to a healthcare organization or a prominent research person publicly, and have others view and benefit from the answer as well.

Given the above, our research focused on data shared via the Twitter platform to assess the level of social awareness around a rare disease such as "Idiopathic Pulmonary Fibrosis" also referred to as "IPF" in the medical field, whose prevalence is researched by Lee et al (2015). For benchmarking purposes, we compared the level of social awareness of IPF with that of a more well-known disease such as "Breast Cancer" that the public hears about more regularly.

Our study revolved around the creation of an “awareness quotient” that quantified the social awareness level of both these diseases, and analyzed if there existed a relationship between the level of awareness triggered by social media, and the number of clinical studies that include both observational and interventional clinical trials conducted by the US National Institutes of Health for that disease. Furthermore, we also study if the trend is progressive from year to year, taking into consideration data subsets for the calendar year 2016 and year to date data for the current year, 2017. For this, the timeline data in Twitter was accessed for a handful of key healthcare groups involved heavily with promoting awareness and relief to the public for each of these diseases.

2 LITERATURE REVIEW

In this paper, as mentioned earlier, our focus was on social awareness of rare diseases rather than more well-known diseases. A rare disease is one that is characterized by its prevalence in a smaller part of the general population, according to Orphanet (2012). Drugs that are developed to treat rare diseases are referred to as “Orphan Drugs”. Given this, rare diseases are also referred to as “Orphan diseases”. According to the study report from the US based Institute of Medicine (2010) on accelerating rare diseases research and orphan product development and the European based Orphanet (2017), an online portal for rare diseases and orphan drugs, there exists around seven thousand rare diseases to date. The working philosophy of these institutions is that though rare diseases may affect a smaller part of the population, there are still too many patients afflicted by rare diseases.

According to Lee et al (2014), for a rare disease such as IPF, though the incidence and prevalence of IPF appear to be increasing, other than lung transplantation, this disease with unknown causes and hence the word “idiopathic” in its name, has no cure and at best, has a dismal 2 to 3-year median survival. Lee et al (2014) use IPF incidence, prevalence and resource cost information from stats across the globe to present evidence pointing to significant symptom and resource burden on IPF patients, most of whom belong to the senior age population.

One of our primary motivations to compare IPF and Breast Cancer in this study is the fact that both of these diseases, though one is rare and the other is not, cause the deaths of approximately the same number of people in the United States annually. In the US, in this year of 2017, around 310,000 people are expected to be diagnosed with Breast Cancer, and around 40,000 will die of it according to stats available from Cancer.Net (2017). Typical breast cancer survival rate is 5-years and early detection is able to extend this to longer periods. For IPF however some 50,000 people are expected to be diagnosed with it each year while 40,000 dies of it every year, according to stats available from the Pulmonary Fibrosis Foundation (2016). As expected, due to the lack of a cure, IPF has an extremely poor prognosis. The question these numbers bring to mind is whether increased awareness about the lack of a prognosis and the abnormally high ratio of IPF’s number of fatalities to the number diagnosed (80%) annually could lead to more clinical studies.

A conceptual study by Muhammad (2017), describes how social media can be effective in highlighting and managing rare long-term health condition such as Alports Syndrome in young people. Muhammad (2017) focuses on various data collection methods through active online participation via social media, and explains how this data can greatly benefit further research and positively impact the disease prognosis.

In the context of this paper, we also collected data from Twitter, but not by eliciting active participation from users. Furthermore, we leveraged the STEPPS (social currency, triggers, emotion, practical value, public, and stories) framework originally created by Berger (2013) and subsequently utilized by Pressgrove et al (2017), in the content of the data collected via Twitter. This was done to tell us if the advocacy groups working on behalf of rare diseases such as IPF could be using any of the attributes of the STEPPS framework to promote more public awareness. And again, as health awareness usually generates more public interest and attention, this could translate to a sufficiently high number of studies and additional research, which could in turn potentially lead us to a cure sooner than later. Pressgrove et al (2017) used a more comprehensive dataset using the Sysomos Media Analytics Platform instead of just a subset of data to search for content related to each attribute of the STEPPS framework. Using the ALS bucket challenge in social media as the case study, they concluded through empirical methods that social currency, positive emotions and public were the most predominant attributes for the ALS social media challenge tweets that went viral. In this paper, we will apply empirical methods similar to what Pressgrove et al (2017) used to identify which tweets and what number could have contributed to a higher level of awareness of IPF.

Similar to Pressgrove et al (2017), another study by Krasnova et al (2008) describes an alternate empirical model that identifies social media participation drivers with respect to needs and peer pressure. In the context of this paper and healthcare in general, since needs mostly always dominate over peer pressure, one approach would be to apply a concept similar to the “needs” part of the model described by Krasnova et al (2008) to compute awareness.

However, in comparing the approaches described by Pressgrove et al (2017) and Krasnova et al (2008), for this paper we focused on the approach described by Pressgrove et al (2017) since (a) it is more objective and (b) relevant, being a recent publication, and as a result will prove to be more effective.

In another study by Hee et al (2017), the conclusion was that the lower prevalence of rare diseases could lead to smaller sample sizes in interventional clinical trials of rare diseases. Hee et al (2017) used empirical methods employing data from clinicaltrials.gov for rare diseases and studied their sample sizes to reach this conclusion. But it is still unclear from their research if smaller sizes in clinical trials impact the overarching intent and accuracy of the clinical study or not. In this paper, we will not be considering the sample sizes of the interventional clinical trials that are part of the clinical studies. The number of clinical studies we refer to in this paper are also obtained from clinicaltrials.gov (operated by the US National Institutes of Health), which include both observational studies and interventional clinical trials.

While various research papers existed on different topics related to this paper, the intent of this paper is to address the following gap, which is to assess if there is a direct correlation between public awareness of a rare disease and the number of clinical studies for it. The emergence of orphan drugs for rare diseases, as we know is primarily driven through clinical studies, and the research paper that comes closest to discussing social media and orphan drugs together is by Milne et al (2017). This paper is a conceptual study of the various approaches involving social media that can further awareness of diseases and how this enables social media to play a role in clinical trials. Nowhere in this study is a *quantitative* assessment of the social media awareness performed using empirical methods, and that is the area or gap which this research paper will be able to fill.

3 DATASET DESCRIPTION

Here, we want to study the correlation between public awareness of a rare disease such as IPF and the number of clinical studies for it during the years of 2016 and 2017. We also want to compare the numbers for IPF with Breast Cancer to gauge the level of awareness and benchmark IPF with respect to a more well-known disease. Given this, we have two sides to our dataset.

1. We used the actual number of clinical ongoing studies for IPF and Breast Cancer in the years 2016 and 2017. For 2017, only the first 3 quarters will be considered. In other words, tweets from 1/1/2017 to 9/30/2017. Here are those numbers from the Clinicaltrials.gov (2017) website, using their “Advanced Search” feature:

Table 1: Number of Clinical Studies by Disease and Year

Disease Name	# Clinical Ongoing Studies	Year 2016 (1/1/2016 - 12/31/2016)	Year 2017 (1/1/2017 - 9/30/2017)
IPF	US	20	61
	World	54	115
Breast Cancer	US	668	1691
	World	1690	2813

2. To gauge the level of social awareness, we obtained data posted in Twitter. The scope of the tweet data collected and processed was limited to the years 2016 and 2017. For 2017, only the first 3 quarters were considered, as we had done the same for the clinical studies’ data. In other words, tweets created from 1/1/2016 to 9/30/2017 were only considered.

3.1 The Twitter API

The Twitter’s Application Programming Interface (API) had the following constraints:

I. When searching for tweets using hashtags, the Twitter system only allowed the data for the past 7 days to be returned. As a result, the Twitter API that we used for programmatic retrieval of tweets using hashtags enforced the same constraint.

II. Secondly, the Twitter system allowed the download of approximately 3,200 latest tweets from a user’s timeline. If data prior to the 3,200 latest tweets was required as in the case of one account or user name (which will be explained later), a web-scraping technique was employed.

For implementation tasks associated with the data collection, the following two python packages were relied on:

1. The python “tweepy” library, which is an easy to use Python library for accessing the Twitter API, and accessible at - http://www.tweepy.org/.
   - This package was used to programmatically download tweets for the different twitter user names.
2. The python “twitter” library, which is a minimalistic Twitter API for Python, and accessible at https://pypi.python.org/pypi/twitter.
   - The “username” search API in this package was used to search for users whose names contained the word “IPF” or “Breast Cancer”.

3.2 Identification of prominent Twitter user names for both IPF and Breast Cancer

The first step in the data collection process involved the determination of which Twitter usernames would most accurately represent communities actively involved in promoting relief and education around IPF and Breast Cancer. Only if this determination occurred, would we be able to download the timeline data (or tweets) from these users’ timelines to do further research on the tweet content.

A first option was to use some prominent hashtags for the two diseases, and search for usernames whose tweets contained these hashtags. For example, use the hashtags “#IPF” and “#BreastCancer” to determine IPF and Breast Cancer related usernames, respectively. An initial script using the “tweepy” python package was created and it successfully downloaded tweets containing these hashtags for the past 7 days. As a next step, the plan was to extract the usernames of the tweet originators and process their timeline data. However, after executing the earlier mentioned script and viewing the actual results, the usernames obtained were few and far in between. In particular, some prominent communities for both of these diseases who had a lot of followers in Twitter were missing from the results if they were inactive for the past 7 days. This would be a serious gap, especially for IPF, in order to have an exhaustive dataset for the implementation.

Given the above gap, this second approach to determine user names proved more effective. The “twitter” python package that supports searching for user names using a keyword was employed.

For IPF: A python routine was created that employed the above package and first searched for the first 100 usernames using the keyword "IPF". Interestingly the acronym "IPF" also stands for some other organizations globally, so some custom coding was required to filter out usernames unrelated to the disease "IPF". In addition, as the word IPF seems to be a part of communities and user names dealing exclusively with IPF, the final list of usernames that fit the filter criteria are:

@EU_IPFF with 498 followers, and 517 tweets
@IPFawareness with 288 followers, and 92 tweets
@FightIPF with 6642 followers, and 434 tweets
@IPFCatalyst with 375 followers, and 480 tweets
@IPFWORLD with 523 followers, and 1104 tweets
@IPFWarrior with 490 followers, and 912 tweets

Using another similar python routine to search for communities that dealt with lung-related issues, and sometimes had "PF" in their description or names, meant that we could get a list of usernames for organizations that may be aware of IPF, but not exclusively dedicated to its cause. This list provided the following results:

@LungAssociation with 38404 followers, and 10128 tweets
@atscommunity with 16277 followers, and 11143 tweets
@EuropeanLung with 7004 followers, and 2959 tweets
@ActionPFcharity with 1602 followers, and 1758 tweets
@patientMpower with 519 followers, and 843 tweets

As the number of usernames dedicated to IPF was just 6, the 5 usernames listed later, though not exclusive to IPF was still included with the notion that we can filter tweets related to "IPF" related words or tags from within their timelines.

For Breast Cancer: A python routine was created that employed the above package and first searched for the first 100 usernames using the keyword "Breast Cancer". The search easily returned 100 usernames and the usernames were sorted in descending order of their number of followers. More public awareness in social media is tied to more followers, hence this rule was observed here as well. Another filter criterion ensured that there was activity within these usernames (or tweets for the years 2016 and 2017 for which we were collecting data). As we had 11 usernames for IPF, the final list of 11 usernames that fit the criteria for "Breast Cancer" were:

@BCCare with 156804 followers, and 33399 tweets
@breastcancernow with 43140 followers, and 50352 tweets
@TheBreastCancer followers, and 35268 followers, and 10066
@BCAction with 23916 followers, and 6829 tweets
@breastcancer with 21034 followers, and 1945 tweets
@PBCC with 20064 followers, and 3492 tweets
@Breastcancerorg with 15729 followers, and 5077 tweets
@BreastCancerH with 14593 followers, and 10883 tweets
@Bakes4Bc with 14041 followers, and 30674 tweets
@BreastYoga with 12157 followers, and 14330 tweets
@thepinkribbon with 11823 followers, and 3545 tweets

For general healthcare: To gauge how both diseases' awareness fared against each other, some general healthcare communities such as the following were included.

@medpagetoday with 54829 followers, and 30429 tweets
@CDCgov with 862739 followers, and 19914 tweets
@AmerMedicalAssn with 648604 followers, and 19807 tweets

The goal was to filter the timeline data of these users and seek out "IPF" and "Breast Cancer" related tweets.

3.3 Download of timeline data for the identified usernames

For IPF: Using the "tweepy" package and the cursor feature that comes with it, the timeline data for the users were downloaded successfully for:

The 6 usernames that were exclusive to IPF (@EU_IPFF, @IPFawareness, @FightIPF, @IPFCatalyst, @IPFWORLD and @IPFWarrior). Since the entire dataset of tweets in each user's timeline was well under 3,200, all of the data was successfully downloaded.

The 3 usernames that were non-exclusive to IPF (@EuropeanLung, @ActionPFcharity and @patientMpower) for whom the entire dataset of tweets in each user's timeline was under 3,200, and hence all of the data was successfully downloaded.

For the two other usernames that were non-exclusive to IPF (@LungAssociation and @atscommunity) for whom the entire dataset of tweets in each user's timeline was over 3,200, a python web-scraping script was employed to download earlier tweets from the timeline from January 1st, 2016. Note that this method did not employ the "got3" python package but simple web-scraping that is possible in the case of these 2 accounts, since their timelines are publicly accessible via an internet browser. This script used a variation of the code publicly available in a github repository and authored by Tom Dickinson (2015).

For Breast Cancer: Using the "tweepy" package and the cursor feature that comes with it, the timeline data for the users was downloaded successfully for all the 11 "BreastCancer" usernames. For the users whose number of tweets exceeded 3,200, the web-scraping method was employed to ensure that data for 2016 was collected for all the user names.

For the general healthcare accounts: Using the "tweepy" package and the cursor feature that comes with it, the timeline data for the users was downloaded successfully for all the 3 usernames (@medpage, @CDCgov and @AmerMedicalAssn). For the users whose number of tweets exceeded 3,200, the web-scraping method developed by Tom Dickinson (2015) was employed to ensure that data for 2016 was collected for all these user names.

3.4 Structure of the downloaded data

The script that downloads tweet data from the user timeline using the "tweepy" package, specifically extracts information from the

"Status" object that represents each downloaded tweet. In particular we extract the "_json" attribute that is an attribute of this object.

Within the "_json" body, though the JSON object has several attributes, the only attributes that we used to extract information from are as follows:

- text (which represents the text of the tweet or the actual tweet content)
- created_at (which represents the time the tweet was generated)
- favorite_count (which represents the number of likes for the tweet)
- retweet_count (which represents the number of times the tweet was retweeted)
- user.screen_name (which represents the username of the user who generated the tweet)
- lang (the language of the tweet)

3.5 Status of the downloaded data

All the tweets for 11 IPF and Breast Cancer related users were initially successfully for the years 2016 and up until the 3rd quarter of 2017 (9/30/2017) for any user using the approaches outlined earlier.

For the general accounts, it was decided that the data collection for these user names beyond the latest 3,200 tweets will be skipped. This is because upon doing a data analysis and inspection of the latest 3,200 downloaded tweets of each of these accounts, no references to IPF were detected, but 124 instances of "breast cancer" related references were detected. Based on this, we could assume and conclude that older tweets in 2016 for these usernames will probably have no references to IPF either, but will definitely have some references to "breast cancer". Given this, it did not make sense to spend any more effort in downloading the earlier tweets as it would have told us nothing new, although this observation was still relevant and only strengthened our hypothesis that IPF is still to this day not discussed at all in the mainstream health related communities, including their social media operations.

In all, we had downloaded 456,412 tweets for the different usernames identified for IPF, Breast Cancer and the 3 general healthcare accounts.

The number of clinical studies that existed for IPF and Breast Cancer was retrieved for each quarter of 2016 and the first 3 quarters of 2017 from the www.clinicaltrials.org website. Numbers retrieved included studies for the US and the entire world, which includes US numbers as well.

Table 2.1: Number of Clinical Studies by Quarter and Year for IPF

Period	Disease	Number of Studies (US)	Number of Studies (World)
1Q2016	IPF	5	10
2Q2016	IPF	2	5
3Q2016	IPF	7	17
4Q2016	IPF	6	22
1Q2017	IPF	8	26
2Q2017	IPF	24	44
3Q2017	IPF	29	45

Table 2.2: Number of Clinical Studies by Quarter and Year for Breast Cancer

Period	Disease	Number of Studies (US)	Number of Studies (World)
1Q2016	Breast Cancer	124	268
2Q2016	Breast Cancer	130	289
3Q2016	Breast Cancer	187	541
4Q2016	Breast Cancer	227	592
1Q2017	Breast Cancer	336	620
2Q2017	Breast Cancer	520	898
3Q2017	Breast Cancer	835	1295

4 RESEARCH DESIGN AND METHODS

Having collected all the tweets for IPF (approximately 15,500), and a good subset of tweets for "Breast Cancer" (approximately 440,800), the immediate objective was to pre-process the obtained data and transform it to a format that would allow us to execute next steps with respect to mining of the content within the tweets.

4.1 Data processing objective

To recap, the objective of the data processing exercise is to answer the question in the research paper, which is to quantitatively determine whether there is a direct correlation between social media tweets and the number of clinical studies ongoing for each disease. In other words, does more public awareness demonstrated by the tweets' characteristics contribute to more clinical studies or not?

To meet the above objective, we created a model that would do the following:

1) Quantitatively calculate social awareness for both diseases, IPF and Breast Cancer, given a collection of tweets for a given period, and whose output is an awareness quotient for the disease.
2) Compare the above quotient calculated using tweets in a given period with the exact number of clinical ongoing studies for the disease, during that same period.

3) Output a final conclusion that indicated how much of a correlation existed or not.

4.2 STEP 1 - Data Pre-processing

As part of this, the entire collection of tweets was run through a preprocessing algorithm that did the following:

1) Since the timeframe of the tweets that we were interested in were those that belonged to 2016 and 2017 (for the first 3 quarters), all tweets whose creation date was outside of this time frame were filtered out and removed.
2) Tweets whose language was not "en" for English were filtered out and removed.
3) For usernames non-exclusive to IPF, non IPF related tweets, identified by whether they contained the words "IPF" or "idiopathic" were filtered out and removed.
4) For the general healthcare accounts, tweets not related to the word "breast" were filtered out and removed.
5) Each remaining tweet was then assigned three additional attributes
   I. A "year" attribute based on the tweet's creation date
   II. A "quarter" attribute based on the tweet's creation date
   III. A "has_link" attribute whose value was set to 1 if the tweet text contained a hyperlink. If not, this value was set to 0
6) Lastly, the tweet text was processed to remove all stop words using the functionality from the nltk.corpus.stopwords python package. The remaining words were also subjected to a lemmatization using functionality from the nltk.stem.WordNetLemmatizer python package.

Following the above, the corpus of tweets was reduced to 444,084. Out of these, 3,492 tweets represented IPF and 440,592 tweets represented Breast Cancer. These tweets were then manually segregated into two separate files which were then used to mine the results for our research.

At this point in the process, for every tweet, we had the following information available:

- Tweet text that was rid of stop words
- Tweet text that was a sentence of "lemmatized" words
- Number of likes for this tweet
- Number of retweets for this tweet
- The year and quarter to which the tweet belongs to
- Whether the tweet contained a hyperlink or not

4.3 STEP 2 - STEPPS classification

This step involved the following:

1) Creation of separate training sets for IPF and Breast Cancer
2) Training separate STEPPS classification models using the Naïve Bayes algorithm for IPF and Breast cancer
3) Execution of the respective models on the entire two datasets for IPF and Breast Cancer such that the entire dataset for each disease was classified into one of the six STEPPS categories

The training sets for both IPF and Breast Cancer were identified from within the corpus of data separately available for each disease.

As described by Pressgrove et al (2017), our plan was to apply the STEPPS framework on this dataset. The STEPPS framework has six categories within it. These are social currency, triggers, emotion, practical value, public, and stories. According to Pressgrove et al (2017), tweets that generally have a high value for the social currency, emotion or public categories were bound to receive the most attention. So, in our case as well, we borrowed a similar principle, but instead of calculating a value for each category within a tweet, we classified the entire tweet into one of the different STEPPS categories.

To be clear, each letter in the "STEPPS" acronym represents a category, which was interpreted as follows, for our purpose:

The ***first*** letter in the "STEPPS" acronym, "S", which stands for "Social currency", represented by "SC" was assigned to tweets that had interesting statistics or eye-catching information. For instance, comparing the number of mortalities of IPF with Breast Cancer indicating percentages is an interesting fact that people can easily understand. These types of tweets with numbers in them were quicker to snag a reader's attention than some others.

The ***second*** letter in the "STEPPS" acronym, "T", which stands for "Triggers" represented by "TR" was assigned to tweets from people who suddenly became aware of the disease due to themselves or a loved one being diagnosed with it.

The ***third*** letter in the "STEPPS" acronym, "E", which stands for "Emotions" represented by "EM" was assigned to tweets from patients who tweet about their personal suffering from the disease.

The ***fourth*** letter in the "STEPPS" acronym, "P", which stands for "Public" represented by "PU" was assigned to tweets about events and initiatives to improve public awareness and get more people involved.

The ***fifth*** letter in the "STEPPS" acronym, "P", which stands for "Practical value" represented by "PV" was assigned to tweets containing information about disease diagnosis and prognosis.

The ***sixth*** letter in the "STEPPS" acronym, "S", which stands for "Stories" represented by "ST" was assigned to tweets from people whose loved ones may have succumbed to the disease. In other words, these people have been impacted personally and have suffered emotional impact due to the disease and the loss of their loved ones.

Given the above, our first goal was to assign one of the STEPPS categories to each tweet within a sample of tweets. To do this we took 110 tweets each from the IPF and the Breast Cancer datasets (which was our training datasets), and manually classified each tweet into one of the STEPPS categories. This way we will had two training sets, one for IPF and the other for Breast Cancer. These training sets were then used to train separate Naïve Bayes classifiers for IPF and Breast Cancer. Once the trained classifiers were available, these were then used to classify the remaining corpus of all the tweets for IPF and Breast Cancer.

At this point, for every tweet, we had the following attributes available:

- Tweet text that was rid of stop words
- Tweet text that was a sentence of "lemmatized" words
- Number of likes for this tweet
- Number of retweets for this tweet
- The year and quarter to which the tweet belongs to
- Whether the tweet contained a hyperlink or not
- The STEPPS category the tweet belonged to

4.4 STEP 3 - Computing Aggregates

In this step, we aggregated the results from Step 2, by year, quarter and disease. In other words, we computed the following information for each year and quarter and disease:

- Total number of tweets
- Total number of likes
- Average number of likes
- Total number of retweets
- Average number of retweets
- Number of tweets in each of the 6 STEPPS categories

(Note: there were 6 metrics obtained in this step)

- Percentage of tweets in each of the 6 STEPPS categories

(Note: there were 6 metrics obtained in this step)

Displayed in **Figure 1**, **Figure 2, Figure 3** and **Figure 4** are plots of these aggregated stats using the python "matplotlib.pyplot" package generated for IPF and Breast Cancer.

Key Observations from the plot in **Figure 1** are as follows:

1) The number of clinical studies are steadily increasing for IPF from quarter to quarter.
2) The percentage of IPF tweets belonging to the "Public" STEPPS category definitely seems to be the dominator among all the STEPPS categories.
3) Average number of likes and retweets are increasing though by a very small value.

Key Observations from the plot in **Figure 2** are as follows:

1) The number of total number of retweets is steadily increasing for IPF from quarter to quarter. A similar increasing trend is observed for the total number of likes and the total number of tweets as well.
2) The number of tweets in the "Public" STEPPS category definitely seems to be the dominator among all the STEPPS categories here as well.

Key Observations from the plot in **Figure 3** are as follows:

1) The number of clinical studies are steadily increasing for Breast Cancer from quarter to quarter. Roughly the number of clinical studies for Breast Cancer is 25 times the number of studies that exist for IPF in a given quarter.
2) On an average, the line plots for each of the aggregated metrics seem to have a constant value. This could be due to the fact that breast cancer has probably reached its peak awareness levels and therefore the trend is holding steady. Even from our data collection exercise, notice that even after pre-processing, we ended up with some 3,000 IPF tweets and 440,000 Breast Cancer related tweets. So again, roughly for every IPF tweet there are roughly 150 Breast Cancer related tweets in existence.

The key observation from the plot in **Figure 4** is that during the 4th quarter of 2016, there was an enormous volume of tweets in the "Breast Cancer" social media world of Twitter. Now even if we consider this quarter as an outlier, generally speaking when it comes to actual numbers instead of percentages, the trend is still upwards with respect to the total number of tweets, likes and retweets for Breast Cancer.

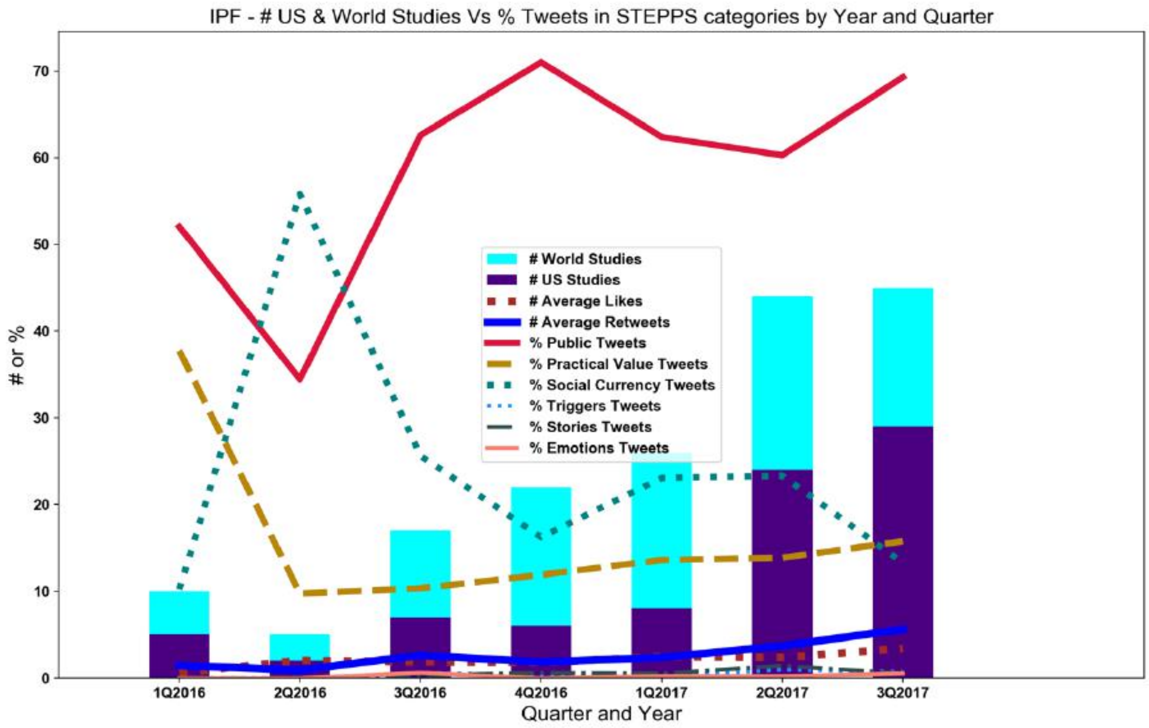

Figure 1: IPF – Number of US and World Studies versus the percentage of tweets in each of the STEPPS categories by Year and Quarter

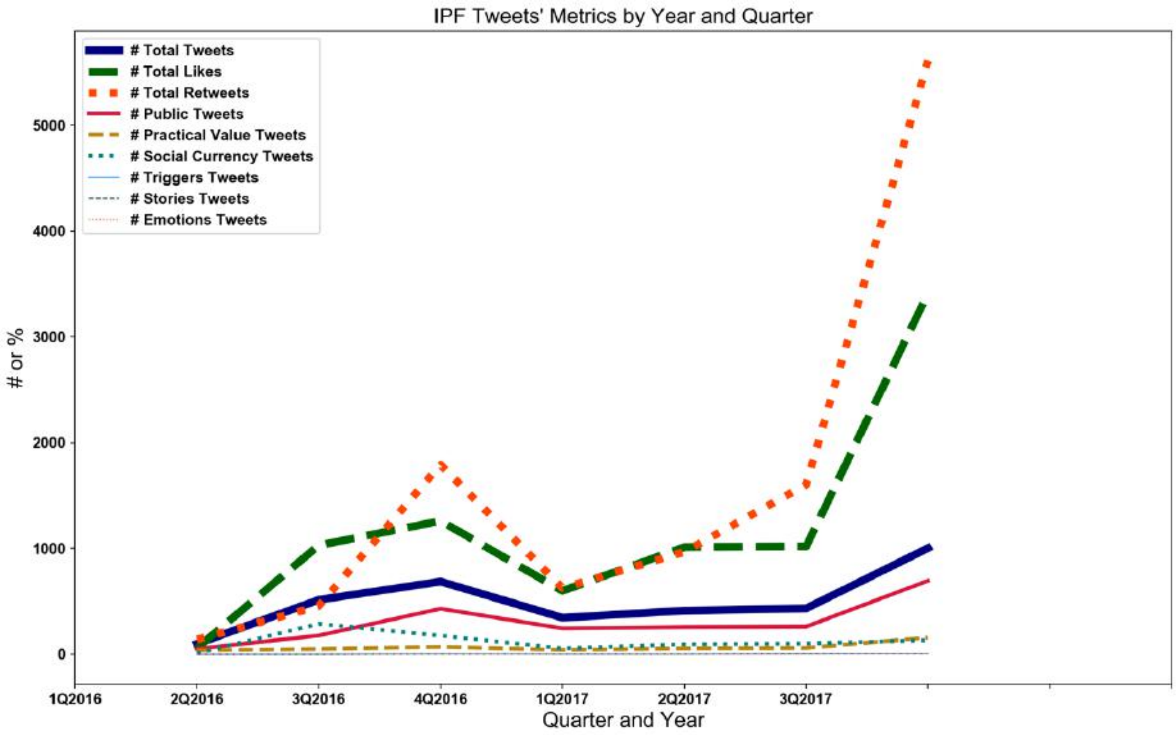

Figure 2: IPF Tweets' Metrics by Year and Quarter

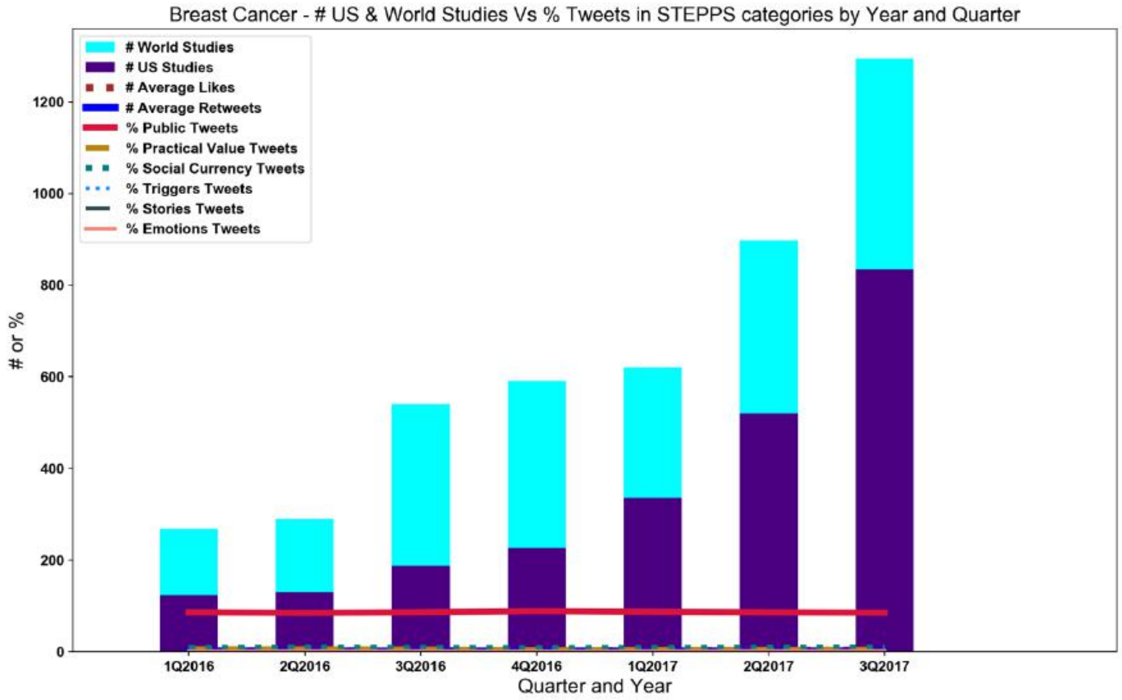

Figure 3: Breast Cancer – Number of US and World Studies versus the percentage of tweets in each of the STEPPS categories by Year and Quarter

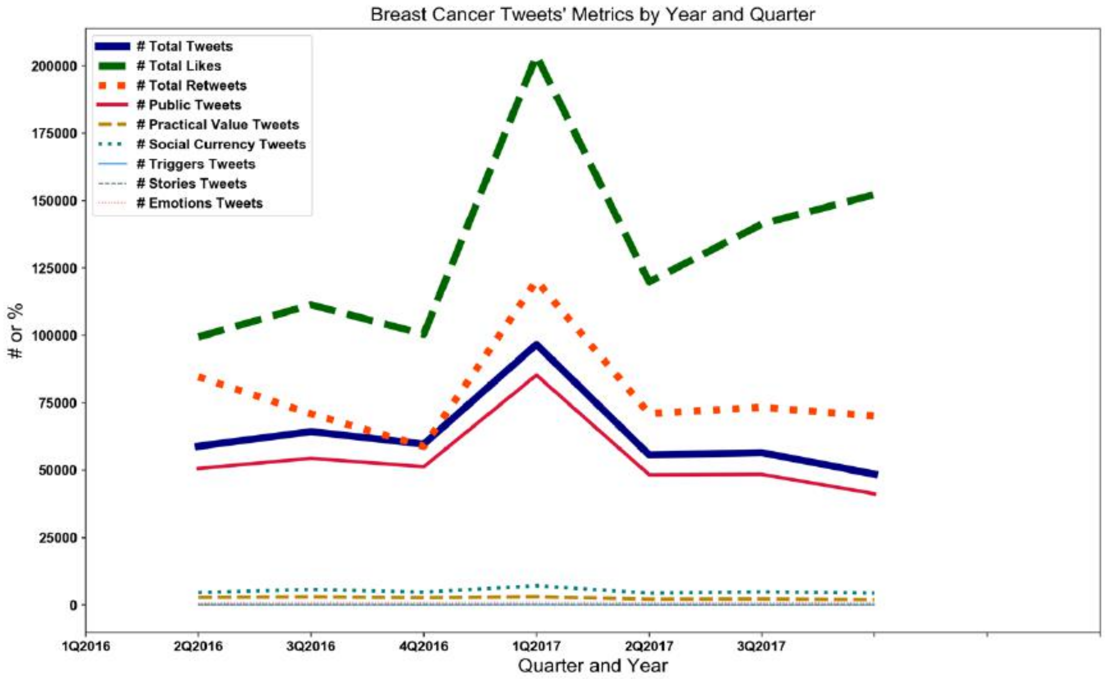

Figure 4: Breast Cancer Tweets' Metrics by Year and Quarter

4.5 STEP 4 - Is there a correlation between the aggregated metrics and the number of clinical studies?

Approach

In order to answer this question, we needed to compute the correlation coefficients of the aggregated metrics to the number of clinical studies.

The aggregated metrics considered were:

- Total number of tweets
- Total number of likes
- Total number of retweets

Note: The next set of metrics will be the total number of tweets in each of the 6 STEPPS categories. As there were 6 metrics obtained for each category, we will list each metric separately for determining correlation

- Total number of tweets in the Social Currency STEPPS category
- Total number of tweets in the Triggers STEPPS category
- Total number of tweets in the Emotions STEPPS category
- Total number of tweets in the Public STEPPS category
- Total number of tweets in the Practical Value STEPPS category
- Total number of tweets in the Stories STEPPS category

We intentionally leave out average and percent computations that are derived from the other metrics, particularly the sum of tweets in different categories and so on. To keep the correlation study simple and rely on metrics that are derived from the raw data, we considered only the totals metrics when trying to study any direct correlation to the number of clinical studies.

As such, we ended up with 9 metrics.

Implementation

The python “statsmodel” package offers functions that allow us to run linear regression also known as the “Ordinary Least Squares” or OLS Regression. We used this method to assess if using the 9 metrics above could be considered as predictor variables and can be used to predict the value of dependent variable which is in our case, the value of the number of clinical studies. In doing so, we obtained a correlation coefficient for each of the 9 metrics or predictors for each disease.

After we determined the correlation coefficients for IPF and Breast Cancer separately, we charted them both on the same plot to view and compare these coefficients.

Comparison

Shown next is the chart where we perform this comparison. However, it is important to keep in mind that the ratio of the number of tweets for Breast Cancer to the number of tweets for IPF in a given month or quarter is very high. In other words, we are talking about a 150:1 ratio. Given this fact, plotting them both in the same scale will result in the IPF data points to overlap the x-axis. This is also why we have not tried to plot IPF and Breast Cancer stats on the same plot.

The key observations from the plot in **Figure 5** are as follows:

1) In general, the correlation coefficients of Breast Cancer are higher than that of IPF, except in the total number of likes, retweets and the number of tweets belonging to the social currency STEPPS category.

2) One reason just the total number of tweets is not having a high correlation could be the fact that by itself the number is not increasing at a rate over each quarter wherein we can see a significant correlation.

3) Though the volume of tweets for IPF is much smaller, the one thing this chart says is that the number of likes and retweets have a significant correlation to the number of studies for IPF and this may be a fact the IPF social media community should leverage more.

4) The key factors influencing Breast Cancer are tweets that belong to the Emotions, Public and Practical Value categories. More of these types of tweets is definitely having a positive and significant correlation to the number of studies for Breast Cancer.

4.6 STEP 5 - Computing the Social Awareness Coefficient

The next step involved calculating the social awareness coefficient using the correlation coefficients computed for IPF and Breast Cancer. As described earlier, we had 9 values representing the coefficients for IPF. Similarly, for Breast Cancer as well, we had 9 values representing its correlation coefficients.

Standardizing to a common correlation coefficient

As visualized in the comparison plot from earlier, some of the coefficients from both diseases were negative. So, the first step was to discard these coefficients from the awareness quotient. We needed to create a value that has positive impact only. Secondly, we also notice that for either of the diseases, some coefficients are larger or smaller for either disease for the same metric. For instance, IPF has a coefficient of 1.2 for total retweets whereas Breast Cancer has 0.22 for the same metric. Knowing that each disease has its own “brand” value in social media, in order to be able to compare them at the same level, it was decided to average the positive coefficients of each for every metric and arrive at a common formula that can be then used to compare both the diseases and quantify their awareness quotient.

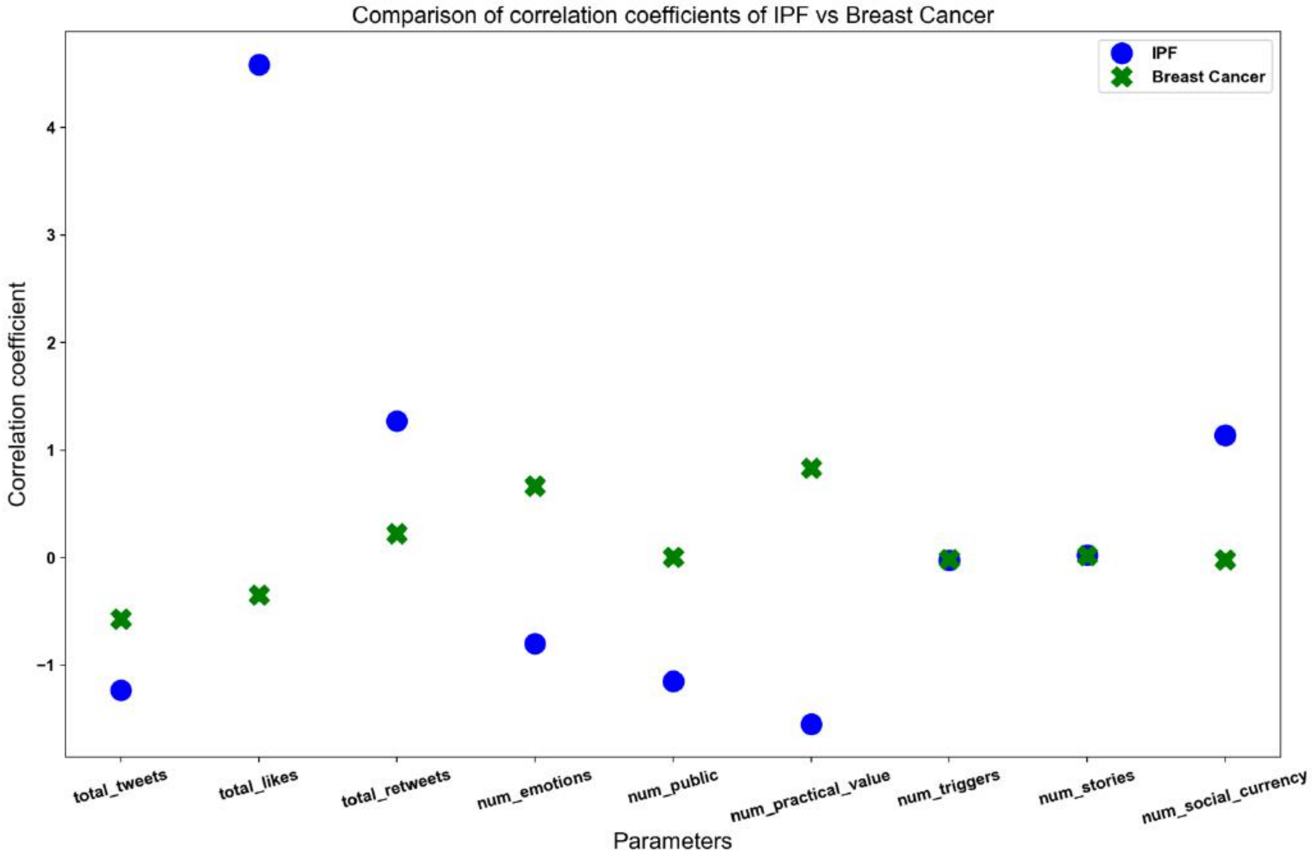

Figure 5: Comparison of Correlation Coefficients of IPF and Breast Cancer

Table 3: Correlation Coefficients by Disease

	Total Tweets	Total Likes	Total Retweets	Number of Tweets in the STEPPS category "Emotions"	Number of Tweets in the STEPPS category "Public"	Number of Tweets in the STEPPS category "Practical Value"	Number of Tweets in the STEPPS category "Triggers"	Number of Tweets in the STEPPS category "Stories"	Number of Tweets in the STEPPS category "Social Currency"
IPF Coefficients Calculated	-1.228	4.585	1.271	-0.796	-1.149	-1.546	-0.021	0.024	1.138
Breast Cancer Coefficients Calculated	-0.568	-0.345	0.224	0.666	0.006	0.832	-0.013	0.021	-0.016
IPF Coefficients After eliminating negative coefficient values	0	4.585	1.271	0	0	0	0	0	1.138
Breast Cancer Coefficients After eliminating negative coefficient values	0	0	0.224	0.666	0	0.832	0	0	0
Average Coefficient Values	**0**	**2.29**	**0.75**	**0.33**	**0**	**0.42**	**0**	**0**	**0.56**

Table 3 shows the calculation that was implemented to arrive at the final coefficients for either disease (Refer to row 5 in the table that has the "Average Coefficient Values").

Final social awareness quotient

After having arrived at a standard or common values of correlation coefficients for either disease, the next task was to create the final formula for the social awareness quotient.

For a given timeframe, be it a year or a month or a quarter, this was obtained by calculating the product of each metric's value for that timeframe by its standard correlation coefficient, and computing the overall sum of all the products, and dividing it by a constant value.

For example, say for a given quarter, the formula would look like this:

Social Awareness Quotient *(for a given timeframe) =*
(Total Number of Tweets Standard Coefficient for Total Tweets+*
*Total Number of Likes * Standard Coefficient for Total Likes +*
*Total Number of Retweets * Standard Coefficient for Total Retweets +*
*Number of "Emotions" Tweets * Standard Coefficient for Number of "Emotions" Tweets +*
*Number of "Public" Tweets * Standard Coefficient for Number of "Public" Tweets +*
*Number of "Practical Value" Tweets * Standard Coefficient for Number of "Practical Value" Tweets +*
*Number of "Triggers" Tweets * Standard Coefficient for Number of "Triggers" Tweets +*
*Number of "Stories" Tweets * Standard Coefficient for Number of "Stories" Tweets +*
*Number of "Social Currency" Tweets * Standard Coefficient for Number of "Social Currency" Tweets)/A constant value*

In the earlier formula, the total number of tweets, retweets and so on considered belong to that given quarter for our calculation.

The reason for choosing a constant is to arrive at a final value whose value is at the same scale as the number of studies. If the final value were too large or too small, it would be harder to ascertain how close this value is to the actual number of studies. In our case, after some experimentation, the value arrived at for the constant was 250.

In addition, given that some of the coefficients were zeroes, the final formula that considered only the non-zero coefficients and sets the constant value to 250, would look like the following:

Social Awareness Quotient *(for a given timeframe) =*
*(Total Number of Likes * Standard Coefficient for Total Likes +*
*Total Number of Retweets * Standard Coefficient for Total Retweets +*
*Number of "Emotions" Tweets * Standard Coefficient for Number of "Emotions" Tweets +*
*Number of "Practical Value" Tweets * Standard Coefficient for Number of "Practical Value" Tweets +*
*Number of "Social Currency" Tweets * Standard Coefficient for Number of "Social Currency" Tweets)/250*

Or, using the actual coefficients arrived at from our dataset, the formula is:

Social Awareness Quotient *(for a given timeframe) =*
*(Total Number of Likes * 2.29 +*
*Total Number of Retweets * 0.75 +*
*Number of "Emotions" Tweets * 2.33 +*
*Number of "Practical Value" Tweets * 0.42 +*
*Number of "Social Currency" Tweets * 0.56) / 250*

Using this formula, the social awareness quotient was calculated for both IPF and Breast Cancer. The tables below each show the calculated values computed for each disease. As is evident the value fluctuates and in some cases. It has a very high value in a certain quarter after which a downward trend is observed in the next quarter.

Given our assumption that social awareness only improved over time, or across each quarter, we also decided to augment the calculated quotients with adjusted values for each quarter, wherein we extrapolated the values between the first and last quarters, using the first and the last quarters as a basis. Having done this for each disease, notice the values in blue in the "Adjusted" columns for each disease.

For our purpose, we are going to consider the adjusted value as the final social awareness quotient for each quarter, accounting for latent factors, outliers and other considerations that impact the social awareness quotient, but were beyond the scope of this project to consider.

Table 4: Calculated and Adjusted Social Awareness Quotients by Disease, Quarter and Year

Quarter and Year	Calculated IPF Social Awareness Quotient	Adjusted-IPF Social Awareness Quotient	Quarter and Year	Calculated Breast Cancer Social Awareness Quotient	Adjusted Breast Cancer Social Awareness Quotient
1Q2016	1.001	1.001	1Q2016	1170.899	1170.899
2Q2016	10.875	8.896	2Q2016	1239.226	1244.025
3Q2016	17.044	16.79	3Q2016	1103.078	1317.151
4Q2016	7.438	24.684	4Q2016	2234.247	1390.277
1Q2017	12.271	32.579	1Q2017	1315.751	1463.403
2Q2017	14.247	40.473	2Q2017	1519.548	1536.529
3Q2017	48.367	48.367	3Q2017	1609.656	1609.656

With the data in **Table 4**, we have the plots in **Figure 6** and **Figure 7** that compare the social awareness quotients for each disease.

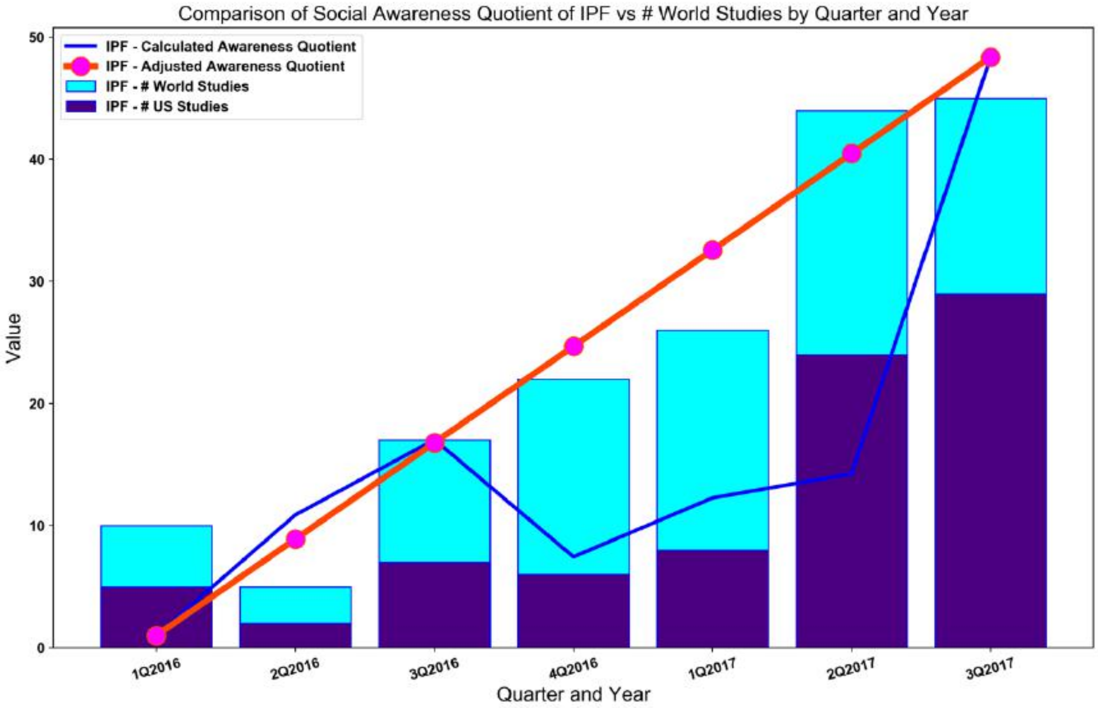

Figure 6: Comparison of Social Awareness Quotient of IPF vs # World Studies by Quarter & Year

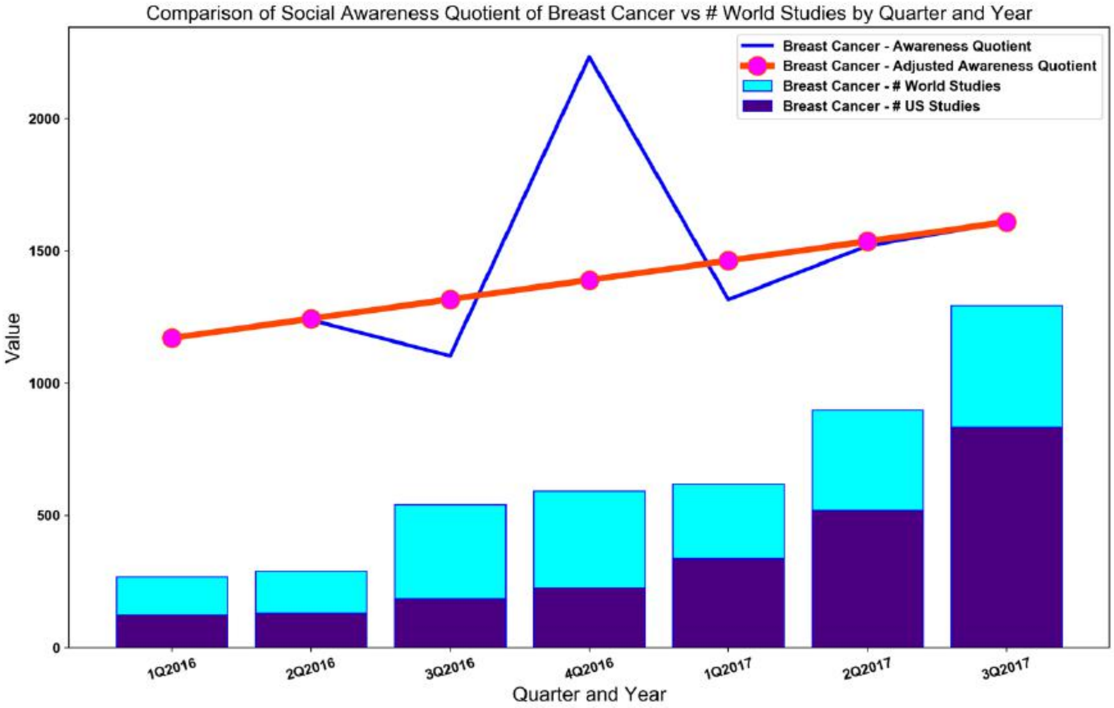

Figure 7: Comparison of Social Awareness Quotient of Breast Cancer vs # World Studies by Quarter & Year

4 CONCLUSIONS

As of the end of 3Q2017, the final social awareness quotient for IPF and Breast Cancer is determined to be around 48 and 1600 respectively. The number of world studies for IPF and Breast Cancer is around 45 and 1295 respectively. To recap, our intent with this original research was to determine if there is a connection or a correlation between the social awareness of a disease (which we calculate using the “tweets” data from Twitter) and the number of clinical studies for that disease obtained from the clinicaltrails.gov website. This correlation was to be measured for each quarter from 2016 January to 2017 September. With the results displayed in the plots in **Figure 6** and **Figure 7**, we can say that there is indeed a close correlation between the two. This is seen by the upward trend in the values of the social awareness quotient and the number of studies for both diseases.

Summary of Results

However, it is to be noted that Breast Cancer’s quotient is still 32 times that of IPF, which is not surprising given what we have been seeing in Twitter in terms of sheer volume of tweets, number of followers, the number of public usernames advocating awareness for Breast Cancer. Drawing a parallel to the number of studies for Breast Cancer, that is 28 times more than that for IPF.

However, given this trend, one could ask when could we expect IPF to achieve a similar social awareness status and have an equivalent number of studies. To answer this, the social awareness quotient for IPF was extrapolated for the next 200 quarters in the next 50 years. The resulting plot is the one displayed in **Figure 8**. From this plot, we can say that knowing what we know today, it will be year 2065 when IPF will enjoy a value of 1,600 as its social awareness quotient. While this is rather depressing given that we may have to wait 50 years to have a better prognosis for IPF similar to what Breast Cancer has today, were earlier detection in place, these are still insights we can share with the advocacy groups who can then work on promoting additional awareness, contributing to more research and ultimately fast-tracking the path to a much better prognosis than the dismal one patients experience today.

A similar plot was charted for Breast Cancer as well in **Figure 9** that displays the fact that the social awareness quotient will be around 16,000, so 10 times what it will be for IPF. This can be explained by the fact that after a certain point in time, the world may be so aware of Breast Cancer that there is not that much of a need to expend social media “currency” on this disease. Also, who knows, there might even be a cure by then, so most of the awareness may be around how to recover from or prevent the disease from occurring in the first place.

Future Research

The above approaches used to determine the social awareness quotient is definitely just one approach. The analysis was performed using tweets from the timelines of usernames that were chosen for this research, hence there is definitely some bias in going with this sample of usernames only. One could probably arrive at a more accurate model were we to repeat several iterations of the above using different sets of usernames and their timeline tweets for each iteration.

In addition, we are not considering any latent factors, outlier reduction and other model tuning opportunities that could further allow us to arrive at a more accurate model. Also, we did have some correlation coefficients turning out to be negative in value. A key metric such as “Total number of tweets” having a negative correlation may be something that needs further investigation. Again, iterating through this approach using a different set of tweets and user names may lead us to better or different insights.

Finally, the total number of followers for each username as well as the number of unique follower names for the entire set of usernames in a given iteration might be useful stats to include in the formula for the calculation of the social awareness quotient.

Last but not the least, the ultimate goal of this paper is to establish a methodical approach to quantitatively measure social awareness, and that has been achieved with the above. Regardless of the approach, it is imperative that social awareness is measured quantitatively especially for rare diseases, so that action can be taken to directly impact their prognosis and treatment be it through clinical studies or orphan drug development and so on. With the approach outlined here, similar models could be developed to bring more social awareness to other rare diseases such as Huntington’s Disease and so on.

ACKNOWLEDGEMENTS

As part of my research, I want to thank pediatric pathologist, Dr. V.G. Chellam, and pulmonologist, Dr. P Arjun, who are both from KIMS (Kerala Institute of Medical Sciences) Hospital, Thiruvananthapuram, from my hometown in Kerala, India, for sharing their insights on the current status of IPF awareness in India. I had lost my own father to IPF 15 years ago and I was curious if anything at all has changed in terms of its prognosis in that part of the world. Unfortunately, the situation remains the same there with life expectancy and prognosis post diagnosis continuing to be dismal as ever, with any public attention and knowledge around this rare disease still remaining at large, even on social media.

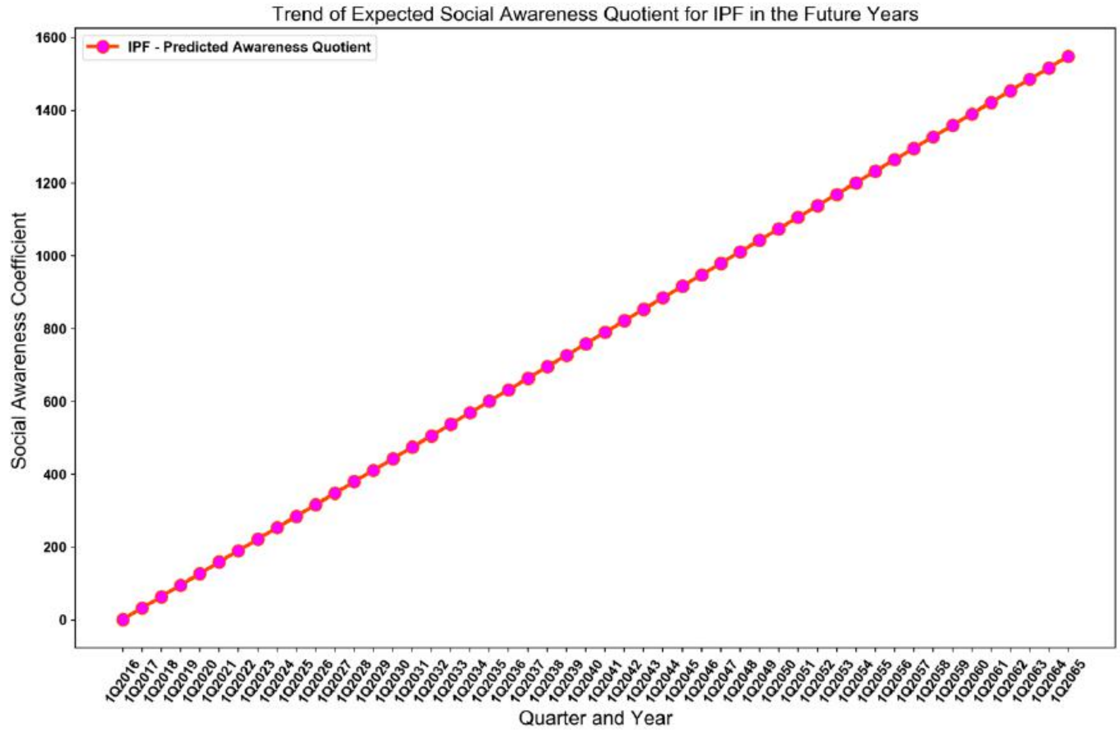

Figure 8: Trend of Expected Social Awareness Quotient for IPF in the Future Years

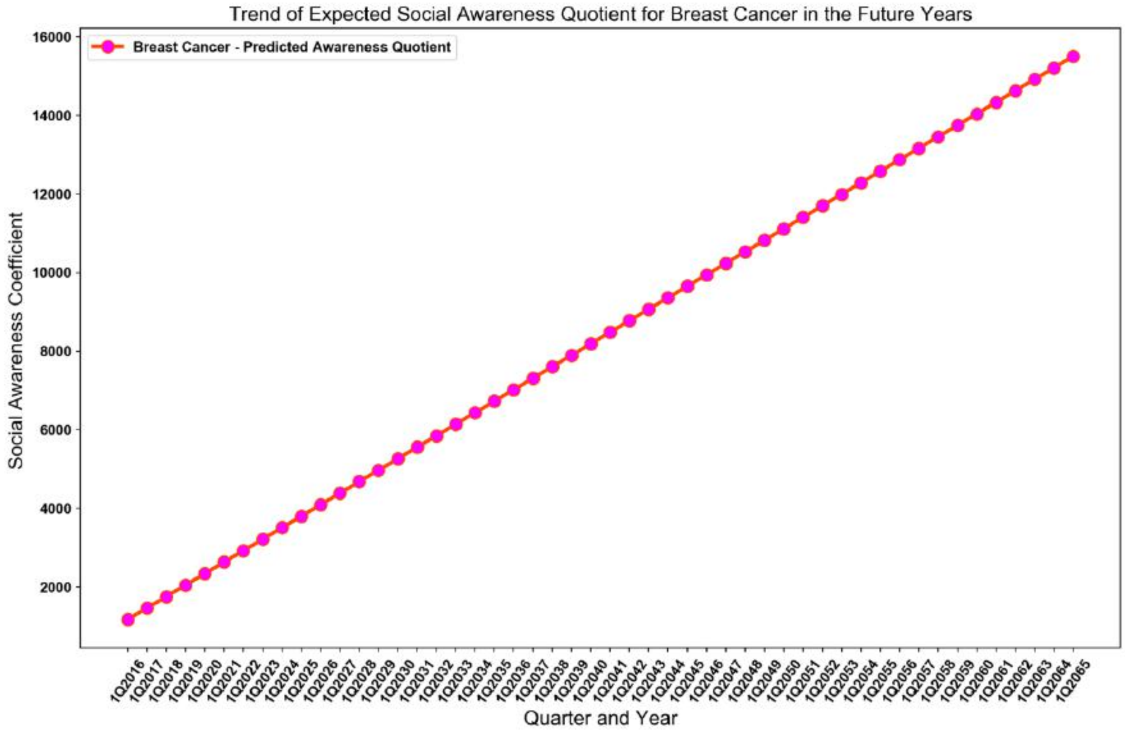

Figure 9: Trend of Expected Social Awareness Quotient for Breast Cancer in the Future Years

REFERENCES

[1] Augustine S. Lee, Isabella Mira-Avendano, Jay H. Ryu, Craig E. Daniels. (2014). “The burden of idiopathic pulmonary fibrosis: An unmet public health need”. Respir Med. 2014 Jul;108(7):955-67. doi: 10.1016/j.rmed.2014.03.015. Epub 2014 Apr 13.

[2] Institute of Medicine. (2010). Rare Diseases and Orphan Products: Accelerating Research and Development (2010). Retrieved on December 2, 2017, from https://doi.org/10.17226/12953.

[3] Hee et al. (2017). “Does the low prevalence affect the sample size of interventional clinical trials of rare diseases? An analysis of data from the aggregate analysis of clinicaltrials.gov”. Orphanet Journal of Rare Diseases (2017) 12:44 DOI 10.1186/s13023-017-0597-1

[4] Pressgrove G, McKeever BW, Jang SM. (2017). “What is Contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge”. Int J Nonprofit Volunt Sect Mark. 2017; e1586. https://doi.org/10.1002/nvsm.1586.

[5] Terence C. Ahern, PhD. (2017). Social Media: Practices, Uses and Global Impact (Chapter 10: Using Social Media to Highlight and Manage Rare and Long-Term Health Conditions in Young People by Shahid Nazir Muhammad, The Renal Patient Support Group (RPSG), England, UK). Nova Science Publishers, Inc. (Pub. Date: 2017 - 4th quarter; ISBN: 978-1-53612-735-5)

[6] Krasnova, Hanna; Hildebrand, T; Guenther, Oliver; Kovrigin, A; and Nowobilska, A, (2008). Why Participate in an Online Social Network? An Empirical Analysis. ECIS 2008 Proceedings. 33. http://aisel.aisnet.org/ecis2008/33.

[7] Phil Baumann (2009, January 16). 140 Health Care Uses for Twitter. Retrieved on December 2, 2017, from https://philbaumann.com/2009/01/16/140-health-care-uses-for-twitter/.

[8] Orphanet (2012, October 25). About rare diseases. Retrieved on December 2, 2017, from http://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN.

[9] Orphanet (2017). The portal for rare diseases and drugs. Retrieved on December 2, 2017, from http://www.orpha.net/consor/cgi-bin/index.php?lng=EN.

[10] Cancer.Net (2017, April). Breast Cancer Statistics. Retrieved on December 2, 2017, from http://www.cancer.net/cancer-types/breast-cancer/statistics.

[11] Pulmonary Fibrosis Foundation (2016, November). About Pulmonary Fibrosis, Symptoms Causes and its Treatments, along with a summary of the Pulmonary Fibrosis Foundation. Retrieved on December 2, 2017, from http://www.pulmonaryfibrosis.org/docs/default-source/marketing-brochures/about-pf_nov2016.pdf.

[12] Sysomos (2017). The Sysomos Media Analytics Platform home page. - Retrieved on December 2, 2017, from https://sysomos.com.

[13] Berger, J. (2013). Contagious: Why things catch on. London: Simon & Schuster.

[14] Clinicaltrials.gov (2017, September). Clinicaltrials.gov website from the US National Institutes of Health. Advanced Search form retrieved on December 24, 2017, from https://clinicaltrials.gov/ct2/results/refine.

[15] Milne, Christopher-Paul et al. (2017). The Use of Social Media in Orphan Drug Development. Clinical Therapeutics, Volume 39, Issue 11, 2173 - 2180. DOI: http://dx.doi.org/10.1016/j.clinthera.2017.08.016.

[16] Stanford University (2009, April 7). Article on stemming versus lemmatization, used in natural language processing. Retrieved on December 2, 2017, from https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[17] Tom Dickinson (2015). Tom Dickinson is a software developer who has described a web scraping method by which a publicly available Twitter timeline data for usernames can be retrieved. This capability is featured in his blog - http://tomkdickinson.co.uk/2015/08/twitter-search-example-in-python/. This script, that he developed allows the download of a user’s tweets from their timeline, if the timeline is accessible publicly. This method uses the web-scraping method along with the “twitter” and other libraries in python and the raw code for the script is available at the link - https://raw.githubusercontent.com/tomkdickinson/Twitter-Search-API-Python/master/TwitterScraper.py. Both of the links specified in this reference were retrieved on December 2, 2017.