Characterizing tweet volume and content about common health conditions across Pennsylvania

Abstract
Background: Tweets can provide broad, real time perspectives about health and medical diagnoses that can inform disease surveillance in geographic regions. Less is known however about how much individuals post about common health conditions or what they post about.

Objective: We sought to collect and analyze tweets from one state about high prevalence health conditions and characterize tweet volume and content.

Methods: We collected 408,296,620 tweets originating in Pennsylvania from 2012-2015 and compared the prevalence of 14 common diseases to the frequency of disease mentions on Twitter. We identified and corrected bias induced due to variance in disease term specificity and used the machine learning approach of differential language analysis to determine the content (words and themes) most highly correlated with each disease.

Results: Common disease terms were included in 226,802 tweets. Posts about breast cancer (22.5% messages, 2.4% prevalence) and diabetes (23.1% messages, 17.2% prevalence) were overrepresented on Twitter relative to disease prevalence, while hypertension (9.9% messages, 36.3% prevalence), COPD (0.9% messages, 8.5% prevalence), and heart disease (7.8% messages, 19.4% prevalence) were underrepresented. The content of messages also varied by disease. Personal experience messages accounted for 12% of prostate cancer tweets and 24% of asthma tweets. Awareness themed tweets were more often about breast cancer (23%) than asthma (6%). Tweets about risk factors were more often about heart disease (10%) than lymphoma (2%).

Conclusions: Twitter provides a window into the online visibility of diseases and how the volume of online content about diseases varies by condition. Further, the potential value in tweets is in the rich content they provide about individuals’ perspective about diseases (e.g. personal experiences, awareness, risk factors) that are not otherwise easily captured through traditional surveys or administrative data.

Keywords: twitter; disease prevalence; big data; public health

Introduction
Communities are increasingly identified as a driver of health, yet our ability to track changes in the health of communities has been limited by the nature of community-level data. These data are typically survey-based or derived from administrative
health care claims. In both of these cases, delays in data availability can preclude timely interventions. Social media channels like Twitter offer a new opportunity to track regional health trends by observing health related communication generated by the public and for the public. [1-7]

There is an opportunity to determine how emerging digital data sources are complementary (i.e., social media data has similar findings to traditional health data sources) and augmentative (i.e., social media provides new real-time information about health not available in data collected through traditional means). To better quantify the value added of social media for public health surveillance, needed is an understanding of how much data exists about different health conditions. High prevalence conditions which affect much of a population may be underrepresented online while low prevalence conditions could be discussed more frequently on Twitter. Further it is likely that there are different drivers (e.g. disease morbidity and mortality, celebrity news, acuity, stigma) that may influence the volume of online health conversations.

To better characterize health-related tweet volume and content, we compared the volume of Twitter messages about common diseases with the prevalence of the disease determined from inpatient and outpatient claims. We then characterized the public perception of common diseases by identifying content (words and themes) most frequently associated with each condition.

**Methods**

This was a retrospective analysis of publicly available data about health conditions posted on Twitter in Pennsylvania. This study was approved by the University of Pennsylvania Institutional Review Board.

We collected tweets related to 5 of the top causes of death in the US originating from Pennsylvania. The causes of death were then further divided into subcategories: heart disease (heart disease, hypertension), diabetes, stroke, cancer (breast, skin, lung, lymphoma, leukemia, prostate, pancreatic, ovarian), and chronic lung disease (asthma, COPD).

**Data Sources**

**Twitter Data**

Twitter is a social media platform which allows users to send and receive 140-character messages or “tweets.” Tweets from 2012-2015 were collected via the Twitter Application Programming Interface as described in Preotiuc-Pietro et al. [8] The county of origin of each tweet user was determined and the dataset was filtered to obtain only tweets for users in Pennsylvania. To increase the sample size of tweets from the state, all unique user ID’s were recorded and used to extract timelines (each user’s prior 3200 tweets).
Disease Keywords
The dataset analyzed was filtered for messages containing at least one keyword referencing a disease. The lexica of keywords (Multimedia Appendix 1) for each disease was derived from the Consumer Health Vocabulary [9] and supplemented by the authors of the study.

Tweet Location
All tweets used in this analysis were classified as originating from a county in Pennsylvania. Tweets were mapped to a county using a combination of coordinates and the user provided location field as per the method described in Schwartz et al. [10] For county mapping, we identified if coordinates were present with the tweet. If coordinates were present, these were used to identify the county of origin. For tweets without coordinates, we used the location field provided in the user’s profile to identify the county. When the field contained only a city, or city nickname, it was mapped to a county as long as it met the following criteria: at least 90% of the population in all the cities with that name are in one specific city. For example, "Chicago" would get mapped to Chicago, Illinois (IL) since greater than 90% of the population in all cities named “Chicago” in the US are located in Chicago, IL. "Springfield" wouldn’t be mapped as there are approximately 50 different regions named “Springfield” in the US of similar population density. The same process in the previous step was used if the county name was listed without a specified state. Cities that were among the top 1,000 English or Spanish nouns, verbs, and adjectives were not considered.

Deriving Topics about Individual Diseases
Utilizing all messages from the dataset, 200 topics (i.e. groups of co-occurring words) were generated using the Mallet implementation of Latent Dirichlet Allocation (LDA). The input data for LDA was filtered to remove all disease keywords along with all words used by less than 5% of tweet authors.

The topic distribution of each individual message was then calculated as described in Schwartz et al. [11] The Pearson correlation between topic distribution and a binary label of whether or not the tweet contained the disease mention was calculated. All correlations were corrected for false discovery rate using the Benjamini-Hochberg procedure.

Organizing Topics into Themes
Ten themes were created by clustering the 200 LDA topics using non-negative matrix factorization (NMF) of the LDA topics derived from the messages. We identified the resulting clusters of topics as “themes.” The LDA topics specify the probability of each word given each topic. NMF provides a weighted value indicating how much each topic, and hence each word in each topic, contributes to each theme. Theme distributions for each message were then calculated in the same manner as described previously for the topic distributions, using Bayes rule to compute \( p(\text{theme}|\text{word}) \). The resulting themes were manually labeled as follows: News,
Research, Slang/Popular Culture Reference, Environment, Diagnosis/Survivorship, Treatment, Diet/Prevention, Awareness, Risk Factor, Personal Experience.

Statistical Analysis

Disease Prevalence
Outpatient and inpatient hospitalization claims were retrieved from 2013 and 2014 claims data from the Pennsylvania Health Care Cost Containment Council. Claims corresponding to each disease were identified using the primary and secondary diagnostic codes which were encoded via the corresponding International Classification of Disease 9th edition. The codes pertaining to a specified disease were determined using the grouping provided by Clinical Classification Software developed as part of the Healthcare Cost and Utility Project. [12] Disease prevalence is defined as the number of unique patients in each county that have a claim related to a given disease divided by the total population of the county. The average of those county level prevalences were used as the state prevalence for each disease.

Adjusted Message Counts/Correction Factors
Due to ambiguity in some of the disease lexica, the message counts for each disease need to be scaled to reflect that many uses of terms such as "heart attack" or "stroke" are metaphorical or refer to other subjects such as golf 'stroke'. The scaling is accomplished via a correction factor based on the manual review of tweets by two researchers using the methods outlined in Weeg, et al. [13] To calculate the correction factor for a disease, a sample of 30 tweets for each keyword were sampled. Those tweets were then classified as being a reference to disease or not a reference to a disease. The percentage of tweets, from the sample, pertaining to disease was identified as the correction factor for that keyword, $w_k$. To calculate the corrected message count for a disease, the product of the correction factor, $w_k$, and the number of messages containing that keyword, $n_k$, are summed for all keywords for a single disease.

$$
corrected\ \text{message\ count} = \sum_{k=1}^{K} w_k n_k
$$

The corrected message count was then compared with disease prevalence.

Comparing Tweet Volume to Disease Prevalence in PA
We used summary statistics to compare volume of posts on twitter with disease prevalence in PA for those conditions.

Associating Disease with Themes
The distribution of themes was investigated using two different metrics: the probability of the theme given the disease and the pointwise mutual information (PMI) between the disease and theme. The probability of the theme given the
disease provides insight into the most prevalent topics of conversation for the given disease.

\[
PMI = \log \frac{p(\text{theme}, \text{disease})}{p(\text{theme})p(\text{disease})}
\]

The PMI of disease and theme provides a measure of how often a disease and theme co-occur relative to how often the two would co-occur if independent of one another. This provides insight into theme-disease co-occurrence which may be somewhat rare, but is significantly different from random chance.

Results

Tweet Volume and Disease Prevalence Comparison

Tweet Volume
The initial sample of tweets from Pennsylvania consisted of 408,296,620 tweets. The data was filtered for messages containing disease related language resulting in a dataset containing 226,802 messages. Breast cancer (n=39156), stroke (n=53858), and diabetes (n=41615) were the most frequent conditions represented in the dataset (Table 1).
Table 1: Characteristics of the study sample: tweet data and user data

<table>
<thead>
<tr>
<th>Disease</th>
<th>Message Count</th>
<th>Correction Factor</th>
<th>Corrected Message Count</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cancer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>39169</td>
<td>100%</td>
<td>39156</td>
<td>19960</td>
</tr>
<tr>
<td>Leukemia</td>
<td>9129</td>
<td>95.1%</td>
<td>8682</td>
<td>5855</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>5745</td>
<td>92.6%</td>
<td>5317</td>
<td>3719</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>5276</td>
<td>93.4%</td>
<td>4927</td>
<td>2758</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>3063</td>
<td>99.9%</td>
<td>3060</td>
<td>1212</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>3231</td>
<td>100%</td>
<td>3231</td>
<td>1189</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>4487</td>
<td>100%</td>
<td>4487</td>
<td>2311</td>
</tr>
<tr>
<td>Skin Cancer</td>
<td>7866</td>
<td>99.9%</td>
<td>7859</td>
<td>4048</td>
</tr>
<tr>
<td><strong>Chronic Lung Disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPD</td>
<td>2137</td>
<td>77.1%</td>
<td>1648</td>
<td>726</td>
</tr>
<tr>
<td>Asthma</td>
<td>18082</td>
<td>92.6%</td>
<td>16742</td>
<td>10185</td>
</tr>
<tr>
<td><strong>Heart Disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>53858</td>
<td>15.1%</td>
<td>8141</td>
<td>34298</td>
</tr>
<tr>
<td><strong>Additional Health Conditions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>41615</td>
<td>96.6%</td>
<td>40217</td>
<td>16321</td>
</tr>
<tr>
<td>Hypertension</td>
<td>18404</td>
<td>93.7%</td>
<td>17245</td>
<td>12203</td>
</tr>
</tbody>
</table>

**Correction Factors and Corrected Message Counts**
Of the 14 diseases, we identified only two diseases, COPD and stroke, with a correction factor below 90% (Table 1). Messages containing terms related to pancreatic and ovarian cancer were always a direct reference to the disease. References to stroke were non-medical or references to other health topics, such as heat stroke, 85% of the time.

**Comparing Tweet Volume to Disease Prevalence in PA**
When comparing prevalence to corrected message counts we identified that hypertension (9.9% messages, 36.3% prevalence), COPD (0.9% messages, 8.5% prevalence), and heart disease (7.8% messages, 19.4% prevalence) were underrepresented on Twitter. Breast cancer was over represented when comparing corrected message counts and prevalence (22.5% messages, 2.4% prevalence).
Characterizing Tweet Topics about Individual Diseases

We identified the correlation between the topics most prevalent in messages for each disease. All correlations shown in Figure 2 are statistically significant ($P < 0.05$). Topics most correlated with asthma were related to first person accounts of managing the disease (*attack, inhaler*), discomfort associated with the disease (*can’t, breath*), or conditions which pose additional risk (*pollution, mold, dust*) such as allergens. The majority of topics associated with cancer referenced some variety of charity campaign (*pink, ribbon, bracelet*) or awareness effort (*support, awareness, October, pink*). Topics related to stroke were rarely related to cerebrovascular accident, but more often related to other definitions of stroke (*e.g. golf stroke, paint stroke, heat stroke*). Diabetes, heart disease, and hypertension messages were correlated with topics which focused on disease management (*weight loss, insulin, reduce stress,*) and lifestyle choices (*diet, exercise*).
<table>
<thead>
<tr>
<th>Condition</th>
<th>n</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>asthma</td>
<td>18082</td>
<td>0.282</td>
</tr>
<tr>
<td>breast cancer</td>
<td>39169</td>
<td>0.384</td>
</tr>
<tr>
<td>copd</td>
<td>2137</td>
<td>0.072</td>
</tr>
<tr>
<td>diabetes</td>
<td>41615</td>
<td>0.215</td>
</tr>
<tr>
<td>heart disease</td>
<td>14740</td>
<td>0.204</td>
</tr>
<tr>
<td>hypertension</td>
<td>18404</td>
<td>0.27</td>
</tr>
<tr>
<td>leukemia</td>
<td>9129</td>
<td>0.211</td>
</tr>
</tbody>
</table>
Characterizing Tweet Themes across Diseases

Probability of Theme Given Disease
The probability of a topic given the disease provides insight into the most prevalent topics of conversation for a specific disease (Figure 3). We identified messages referencing breast cancer were more likely to be about disease awareness (23%). Heart disease messages mostly focused on *risk factors* such as stress, sleep, and obesity (20%). In most cases asthma messages referenced a personal experience.
Figure 3: Theme Distribution

Pointwise Mutual Information

PMI provides a measure of association between the theme and the disease (Figure 3). We found that diagnosis was a small proportion of the theme distribution for each disease. However, if diagnosis or survivorship is mentioned it is much more likely to be mentioned in conjunction with lymphoma and leukemia than with the other diseases (PMI 0.67 - 0.96). Similarly, a relationship between the risk factors theme and hypertension and heart disease were found (PMI 0.54 - 0.77).
Discussion
There is increasing focus on the potential for big data from digital sources in healthcare. There are challenges associated with using these sources as they are not always collected for the purposes of health tracking.

We explored the potential for using Twitter to better understand the online conversation about common health conditions. We identified that in some cases, traditional health metrics are associated with the volume of tweets for a given disease. While traditional methods of determining disease prevalence are robust they are often delayed in availability as the process for data acquisition and tracking to determine reliable and valid estimates is considerable. Twitter data is available in real time, much faster than traditional methods, and with significant volume providing a measure of public discourse about health. While tweets would not replace traditional surveillance in the way initially posed by Google flu trends [14] they do provide something unique that prevalence statistics do not; a narrative about patient and public thoughts, knowledge, and experiences with health. Twitter provides context to the conversation surrounding disease and allows for characterization of public discussion of high prevalence conditions. We identified that individuals are using Twitter to talk about several diseases although variation exists in the frequency of disease mention and the content.

We observed that people are using Twitter for talking about the most common health conditions in Pennsylvania. Prior work has demonstrated that Twitter has been used to monitor influenza [15] postpartum depression [16], concussion [17], epilepsy [18], and migraine [19]. Prevalence of disease has been correlated with frequency of Twitter posting across a variety of diseases. [13][20] We also identified variability in disease mentions and the specificity of terms. This finding provides us with several insights. First, heart disease and stroke cannot be analyzed without pre-processing due to the ambiguity of many of the keywords associated with the diseases. To resolve these varying issues, other methods will need to be developed to filter out much of the noise associated with these diseases. However, this finding also assures us that the majority of the language we find associated with other diseases can be analyzed using the open vocabulary methods previously described with minimal pre-processing.

Although disease prevalence often coincides with disease mention on Twitter, we found significant variability. The frequency of mention of breast cancer on Twitter were several orders of magnitude higher than lung cancer although lung cancer has a higher rate of death and relatively similar prevalence. Breast cancer has a large social media presence due to awareness and charity campaigns in conjunction with a large community base from those affected by the disease. Lung cancer is tweeted about less often, and is often the result of a pop culture reference from television or a celebrity death.

Traditional metrics provide detailed information about prevalence but not insights
about people's understanding, concerns, questions about health and disease. Our analysis identified several underlying themes that are specific to some diseases. Asthma tweets included references to personal experiences for both the person with asthma as well as parents expressing concern for their children's asthma issues. Although the largest portion of tweets for the different types of cancer analyzed often referenced charity and awareness, we observed that across diseases in our sample, cancer conditions had the largest portion of tweets about diagnosis.

Limitations
We compared data from Twitter 2012-2015 with disease prevalence from 2014 so there may be some variability by year in these estimates. We evaluated unadjusted data from one state so this may not be representative of the conversation about health conditions across other states or geographic regions. Twitter data primarily originates from urban areas, so it may not be the most representative sample across the state of PA. Future work could explore variations in language on Twitter relative to the size of geographic regions, socioeconomic factors (e.g. race, income, urban/rural), and variations in news events or other triggers. Although our correction method eliminates non-disease references it does not account for metaphorical and joking tweets. This impacts diseases such as heart disease, diabetes, and hypertension.

Conclusions
We identified that the volume of tweets is often related to rates of health conditions across a state. The semantic content provided from Twitter provides insight into public perception and awareness of disease beyond what is available through traditional measures of disease prevalence.

Acknowledgements
This project is funded, in part, under a grant with the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Conflicts of Interest
none declared

Abbreviations
LDA: latent Dirichlet allocation
PMI: pointwise mutual information

Multimedia Appendix 1
This study focuses on 14 diseases and each disease is represented by a lexicon of disease related terms. The appendix contains each of the 14 diseases along with the 274 terms which comprise the lexica.
References


