What Do People Tweet When They're Sick? A Preliminary Comparison of Symptom Reports and Twitter Timelines

Ashlynn R. Daughton[1,2] and **Michael J. Paul**[1] and **Rumi Chunara**[3,4]
[1] Department of Information Science, University of Colorado, Boulder, CO, USA
[2] Analytics, Intelligence, and Technology Division, Los Alamos National Laboratory, Los Alamos, NM, USA
[3] Department of Computer Science & Engineering, New York University, New York, NY, USA
[4] College of Global Public Health, New York University, New York, NY, USA
ashlynn.daughton@colorado.edu, michael.paul@colorado.edu, rumi.chunara@nyu.edu

Abstract

Internet data have been shown to be related to disease outbreaks and useful additions to disease surveillance models. However, prior research that has aimed to identify individuals that were experiencing illness suffered from a lack of ground truth. Using data from GoViral, a platform developed to generate self-reported symptoms from a cohort of lay volunteers, we aim to identify when and if Internet data can be used as a reliable indicator of actual illness.

Introduction

Internet data (e.g., search queries, social media posts) have been shown to be predictive of disease outbreaks (Ginsberg et al. 2009; Santillana et al. 2015), though these models have also gotten it wrong (Lazer et al. 2014). Understanding the validity of an internet-based surveillance system, and why it might fail, requires a complete understanding of how the data (internet messages or similar digital artifacts related to illness) relates to the population statistic being estimated (the proportion of people in a population that are ill).

While much has been written about the representativeness of social media (Mislove et al. 2011; Ruths and Pfeffer 2014), a population estimate derived from internet data is not merely a biased sample of the population, which is in some cases a solvable problem (Weeg et al. 2015). There are at least two additional sources of noise which have received less attention in disease surveillance:

- **Construct validity:** Does an internet message about an illness truly represent an instance of illness? Are ambiguous signals being interpreted correctly? Are internet users diagnosing themselves correctly?
- **Missing data:** If someone does not post about being sick, does that mean they are not sick? Is the data missing at random, or are there biases (e.g., demographic) in the self-disclosure process?

Answering these questions requires ground truth data about *individuals*, not merely population-level statistics, which is the standard comparison for validation.

This study compares individual-level health data, including weekly symptom self-reports and viral diagnostic data collected through the GoViral platform, with Twitter messages posted by the individuals.

This work-in-progress paper provides an overview of the dataset, describes key statistics, and outlines our plans for future work.

Methods

This study was approved by the University of Colorado Boulder Institutional Review Board.

GoViral

The GoViral platform was developed to generate self-reported symptoms and bio-specimens from a cohort of lay volunteers. Volunteers were recruited, given a kit (collection materials and customized instructions), instructed to report their symptoms weekly, and when sick with cold or flu-like symptoms, requested to collect bio-specimens (saliva and nasal swab). Symptoms included those common to acute-respiratory infections (fever, cough, sore throat, shortness of breath, chills, fatigue, body aches, headache, nausea, and diarrhea). Bio-specimens were tested for the presence of a panel of acute respiratory infections. Demographic information (age, gender, ethnicity, location), as well as Twitter handle (optional) were also collected from each participant. Full details of the protocol have been reported here: (Goff et al. 2015; Ray and Chunara 2016).

In total, there were 396 participants that shared Twitter handles. Of these, we were unable to obtain Twitter data for 84 (because they had private accounts ($n = 25$), had never tweeted ($n = 4$), or because the Twitter handle provided did not exist on Twitter at the time we collected data ($n = 55$).

For the purposes of the analyses presented here, a 'positive symptom report' is a survey that included any symptom. A 'negative symptom report' are those that indicate the individual was feeling no symptoms at the time of survey submission.

Twitter

For participants that that shared Twitter handles, we used the Tweepy API (Roesslein 2009) to collect Twitter timelines as far back as the Twitter API allows (3,200 tweets per user).

To identify health related tweets, we queried all timelines for tweets that included the following keywords:

- *General words*: 'flu ', ' sick', 'throat', 'hurt', 'sinus', influenza', 'stomach', 'tummy', 'respiratory', 'nose', 'feeling', 'cold', 'feel ', 'h1n1', 'h3n2', 'h5n1', 'flua', 'flub', 'infection', ' ill'
- *Symptoms*: 'fever', 'cough', 'migrane', 'congested', 'stuffy', 'headache', ' ache', 'sore',' head ', 'phlegm', 'sneeze', 'asthma', 'pneumonia'
- *Medications*: 'medicine', 'dayquil', 'nyquil', 'tamiflu', 'mucinex', 'theraflu', 'tylenol', 'motrin', 'aleve',, 'naproxen', 'ibproufen', 'acetaminophen', 'advil', 'virus', 'oseltamivir', 'peramivir', 'infection', 'zanamivir', 'antiviral', 'guaifenesin', 'robitussin', 'phenylephrine', 'decongestant', 'pseudoephedrine', 'antihistamines'

This resulted in 3,188 tweets. We then hand-coded each tweet as relevant or not relevant, where relevant means that the tweet appeared to be an authentic description of the individual feeling poorly, with no other explanation. Mentions of events outside of infectious disease that could account for feeling ill were excluded (e.g., recent surgery, consumption of alcohol, feeling cold because the heater was broken etc.).

This resulted in **266 health related** tweets that could potentially be attributed to seasonal cold or flu viruses.

Results

Our data are summarized in Figure 1. Tweet frequency, labeled health tweets, symptom surveys, and virologic results are shown for each anonymized individual. Overall, 1304 surveys were completed, of which 426 had at least one positive symptom. In addition, there were 17 virologic samples collected, of which 5 were positive for at least one pathogen.

We find a few interesting patterns. First, and perhaps most important, is the observation that tweets that are descriptions of an individual's poor health status that can be attributed to seasonal influenza or other virus are rare. Further, it is even more rare for these tweets to coincide with a positive symptom report.

Among the 266 identified instances of individuals tweeting about feeling ill, only 102 were from individuals that completed at least one symptom report. Of those, only 3 were tweeted within 2 weeks (one week before or one week after) of a positive symptom survey. In contrast, across all users there were a total of 58,408 tweets within 2 weeks of a positive symptom survey. As such, health tweets comprised an extremely small percentage of the tweets written near a positive symptom survey. Overall, in the dataset, users tweet on average 35 times a week (95% confidence interval: 34.6-35.3).

While explicit mentions of illness on Twitter seem rare, we do observe other changes in Twitter behavior during illness. We find some preliminary evidence that individuals tweet less when they are sick. On average, after normalizing to an individuals average number of tweets per week, users tend to tweet 60% less in weeks where they returned a positive symptom survey (95% confidence interval: 59%-61%) compared to no survey. Further, on the exact day that an individual returned a positive symptom survey, users tweet on average only 17% as many tweets (95% confidence interval: 16.7%-17.9%) compared to any other day.

Conclusions

Overall, based on this preliminary data, we find limited evidence that some individuals tweet about their health status. However, we find additional evidence that individuals tweet about symptoms that, for whatever reason, they choose to not report on symptom surveys. It could be that the individuals felt their symptoms did not rise to the level of severity required to report them, they could have forgotten them, or they could have chosen not to disclose those particular symptoms for a variety of reasons.

Future Directions

Although these data are sparse, we plan to further delve into possible differences between those that had infections (either virologically confirmed, or more specific symtpom combinations) compared to others that remained healthy throughout the season.

We also plan to further describe the content of health related tweets. In particular, we aim to find if individuals that tweet about getting sick are more likely to also tweet about other health related topics (e.g., exercise, mental health, health research), and to explore the demographic differences in users who do versus do not share these types of tweets.

Acknowledgements

This work was supported in part by grants from the National Science Foundation (IIS-1643576 and IIS-1551036). LA-UR-18-24424

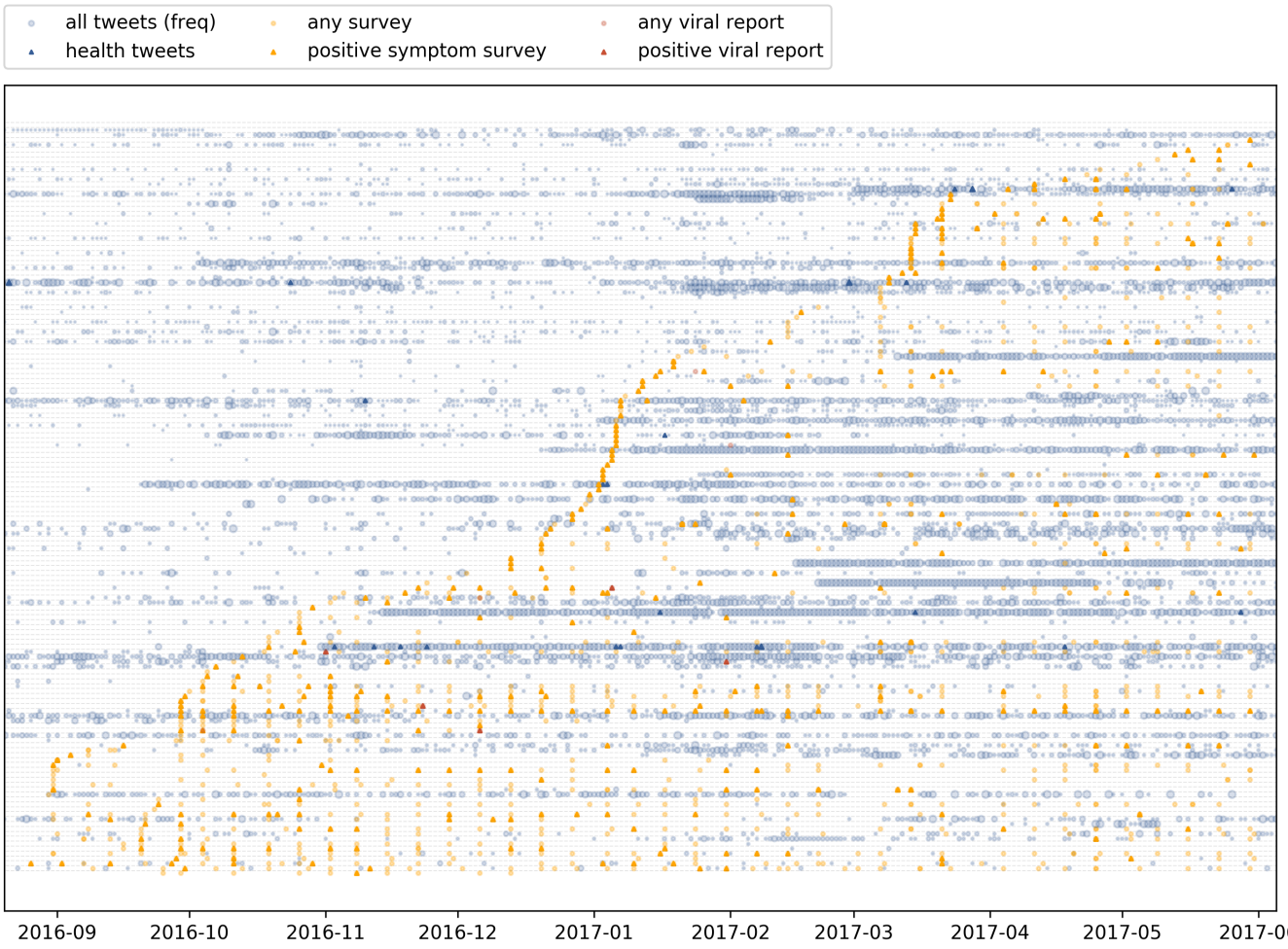

Figure 1: Each row is an individual study participant, sorted by the date of their first GoViral survey. Points correspond to tweets (sized by the number of tweets on that date), survey reports, and diagnostic reports on each date.

References

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Goff, J.; Rowe, A.; Brownstein, J. S.; and Chunara, R. 2015. Surveillance of acute respiratory infections using community-submitted symptoms and specimens for molecular diagnostic testing. *PLoS currents* 7.

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176):1203–1205.

Mislove, A.; Lehmann, S.; Ahn, Y.-y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of Twitter users. In *International Conference on Weblogs and Social Media (ICWSM)*, 554–557.

Ray, B., and Chunara, R. 2016. Predicting acute respiratory infections from participatory data. In *International Society for Disease Surveillance Conference*.

Roesslein, J. 2009. Tweepy Documentation. {http://docs.tweepy.org/en/v3.5.0/}.

Ruths, D., and Pfeffer, J. 2014. Social media for large studies of behavior. *Science* 346(6213):1063–1064.

Santillana, M.; Nguyen, A. T.; Dredze, M.; Paul, M. J.; Nsoesie, E. O.; and Brownstein, J. S. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology*.

Weeg, C.; Schwartz, H. A.; Hill, S.; Merchant, R. M.; Arango, C.; and Ungar, L. 2015. Using Twitter to measure public discussion of diseases: A case study. *JMIR Public Health and Surveillance* 1(1):e6.