# Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness through Twitter

**Michael C. Smith, David A. Broniatowski**
George Washington University
Washington, D.C. 20052
{mikesmith,broniatowski}@gwu.edu

**Michael J. Paul**
University of Colorado
Boulder, CO 80304
mpaul@colorado.edu

**Mark Dredze**
Johns Hopkins University
Baltimore, MD 21211
mdredze@cs.jhu.edu

## Abstract

This study analyzes temporal trends in Twitter data pertaining to both influenza awareness and influenza infection during the 2012–13 influenza season in the US. We make use of classifiers to distinguish tweets that express a personal infection ("sick with the flu") versus a more general awareness ("worried about the flu"). While previous research has focused on estimating prevalence of influenza infection, little is known about trends in public awareness of the disease. Our analysis shows that infection and awareness have very different trends. In contrast to infection trends, awareness trends have little regional variation, and our experiments suggest that public awareness is primarily driven by news media.

## Introduction

Influenza surveillance plays a critical role in the public health mission of agencies at the national and local level. However, infection is only part of the story. Several studies have pointed out that a population's awareness of a disease, and their reaction to it, are major factors that can influence the spread of a disease (Funk et al. 2009; Jones and Salathe 2009; Granell, Gomez, and Arenas 2013). Public health organizations must often manage the perceptions of a population to effectively respond to an outbreak. However, there is no effective, efficient and up-to-date method for tracking population awareness of influenza in the way that there are clinical surveillance systems for influenza infection.

Enter web and social media data. In the past few years, there have been numerous studies demonstrating the ability of these data sources to provide cheap, real-time data for influenza surveillance. Google Flu Trends (GFT) (Ginsberg et al. 2009), among other systems (Yuan et al. 2013; Santillana et al. 2014; Preis and Moat 2014), demonstrated the ability to track flu rates based on search queries. Work using Twitter extended this ability to social media (Culotta 2010; Aramaki, Maskawa, and Morita 2011; Lampos and Cristianini 2012). These data sources are now recognized as real-time enhancements to the existing traditional influenza surveillance infrastructure.

However, during the 2012–13 flu season, which we examine in this study, GFT received widespread criticism

for failing to accurately track the flu (Lazer et al. 2014; Santillana et al. 2014). A primary cause of the inaccuracies was the conflation of search behaviors looking for general information about the flu (awareness) versus those by searchers looking for treatment information (infection). By failing to recognize this distinction, GFT greatly overestimated the infection rate. In contrast, work on Twitter by Lamb, Paul, and Dredze (2013) and Broniatowski, Paul, and Dredze (2013) explicitly modeled this distinction between awareness and infection. This work demonstrated that statistical machine learning classifiers could effectively differentiate awareness from infection. This allowed for the isolation of influenza infection tweets for surveillance, improving over GFT (Paul, Dredze, and Broniatowski 2014).

In this paper, we consider the tweets previously ignored during influenza surveillance: tweets that demonstrate awareness of flu. Twitter provides a solution to the awareness surveillance problem. While those that are concerned about flu in general may not go to their doctor (and are thus not captured by clinical surveillance systems), they are likely to share these concerns with others, including via social media.

This paper presents steps toward understanding and building a real-time surveillance system for disease awareness during an epidemic, specifically that of influenza. The construction of this system allows us to not only provide public health officials with awareness trends, for which they often have no other data source, but also study what drives awareness of influenza in a population. Our analysis will validate our surveillance capabilities—when gold standard data is available—and characterize the surveillance signal to determine which factors most influence a population's awareness.

## Data

This study focuses on the 2012–13 flu season in the United States, from September 30, 2012 (week 40) to May 25, 2013 (week 21). This is a particularly relevant flu season for this study since it had a high peak rate of infection and received considerable media attention. We consider national data as well as regional data for the ten regions defined by the Department of Health and Human Services.[1]

---

[1] http://cdc.gov/flu/weekly/regions2009-2010/hhssenusmap.htm

## Twitter

We downloaded Twitter flu trends from HealthTweets.org (Dredze et al. 2014), which tracks weekly levels of influenza in the United States. The data are based on the state-of-the-art influenza system of Lamb, Paul, and Dredze (2013). This system uses NLP to classify flu-related tweets into two categories: tweets that indicate a personal infection ("I'm sick with the flu") and tweets that express an awareness of influenza ("worried about getting the flu").

Our study makes use of both types of tweets. To estimate the level of influenza in a given week, we count the number of tweets classified as infection-related and we normalize the count by the total number of publicly available tweets published that week. We similarly estimate the level of influenza awareness in a given week by using normalized counts of tweets classified as awareness-related. We restrict our experiments to tweets from the US, using the Carmen geolocation toolkit (Dredze et al. 2013). Full details of the computational pipeline can be found in Broniatowski, Paul, and Dredze (2013).

## Government Influenza Statistics

The Centers for Disease Control and Prevention (CDC) maintains the US Outpatient Influenza-like Illness Surveillance Network (ILINet), a series of healthcare facilities throughout the United States that report weekly cases of influenza-like illness. The CDC releases weekly numbers from the network throughout flu season. The data are publicly available via the CDC's flu dashboard[2] for current and past flu seasons. For this study we obtained the latest available ILINet values for the 2012–13 flu season, using the normalized "weighted ILI" values that adjust for regional differences.

## News Media Volume

We obtained weekly counts for the number of flu-related newspaper articles from NewsLibrary.com, a news archive indexing thousands of US publications. We queried for articles matching the keywords "influenza" or "flu" in the headline or body. We also obtained the US state of the publication associated with each article (if not national), resulting in weekly article counts for each location.

## Analysis

We organize our analysis towards answering the question: what factors influence the public's awareness of influenza? We address this question in multiple stages.

## Comparison to Gold Standard Data

We first compare the Twitter data—both infection and awareness—to the CDC's ILINet data, which is considered the gold standard for measuring prevalence of influenza-like illness in the US. The purpose is twofold. First, it evaluates the ability of the Twitter infection classifier to track influenza prevalence in the ten HHS regions, which are a level of geographic granularity at which the HealthTweets.org

| Region | Infection | Awareness |
|--------|-----------|-----------|
| 1 | .802 | .588 |
| 2 | .804 | .620 |
| 3 | .815 | .575 |
| 4 | .812 | .489 |
| 5 | .818 | .547 |
| 6 | .868 | .633 |
| 7 | .885 | .626 |
| 8 | .869 | .667 |
| 9 | .778 | .548 |
| 10 | .846 | .658 |
| National | .827 | .555 |

Table 1: Correlations between the Twitter infection and awareness data and the CDC's ILINet influenza prevalence data.

data have been validated. Second, and more central to this study, it illustrates how well influenza awareness aligns with influenza prevalence.

Table 1 shows the Pearson correlation between the weekly ILINet data and the two types of Twitter data. At the national level, the awareness data have a significantly lower ($p$=.029) correlation than the infection data, which shows that the distinction between infection and awareness matters for influenza surveillance, and it also shows that awareness does not fully align with prevalence.

## Characteristics of Awareness

We seek to understand the general characteristics of how the weekly awareness levels vary across the season, in contrast to infection levels, which have well-understood trends. Both the Twitter-derived awareness and infection values for all ten regions across the 2012–13 season can be seen in Figure 1(a). Looking at the figure, we make several observations.

First, the regional awareness trends are more similar to each other than the regional infection trends. Awareness rises sharply in week 2 of 2013 in all ten regions, peaking in eight of the ten regions. The infection trends, in contrast, have more variation in their peaks.

We note that the cities of Boston and New York respectively declared public health emergencies in response to the influenza epidemics on January 10[3] and 12,[4] corresponding to week 2 of 2013, which is when most of the awareness trends peaked. News of these epidemics may have driven national awareness.

We quantified the variance of the two trends across regions by computing the Pearson correlation between each region's values and the values from all other regions (that is, the national value if the states from the current region are excluded). Averaged across the ten regions, the infection levels had a mean correlation of .984 (SD .011), while infection levels had mean .963 (SD .036). Thus, each region's trend was generally very similar to all other regions for both awareness and infection, but this was particularly true for

(a) Awareness and infection by region
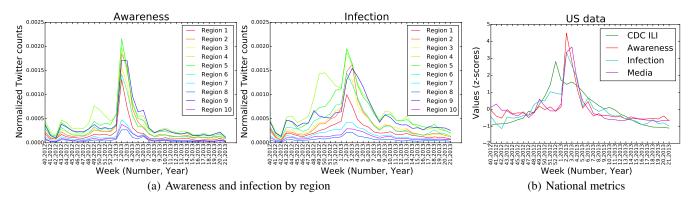
(b) National metrics

Figure 1: The weekly data used in the study, by region (focusing on the two Twitter classifications) and national (all data types).

awareness, which had a higher correlation and lower variance than infection.

Second, awareness has sharper peaks than infection. Infection rises gradually, and drops relatively slowly after the peak. In contrast, the awareness trends rise very sharply, and drop sharply after the peak in most regions. Regions 8 and 9 (the West and Southwest) are exceptions to this, which have a two-week wide peak. One hypothesis is that awareness in these regions rose in tandem with national awareness when flu increased sharply on the east coast, but awareness remained high when the flu spread to the west coast one week later.

Third, outside of the peak, awareness levels are lower than infection levels—at these times people on Twitter don't discuss flu unless they have it. Awareness levels surpass infection levels during the peak, which is believed to be driven by media attention (and a primary cause of Google Flu's overestimate of the peak, as discussed in the Introduction). The effect of news media is discussed further below.

**Effect of News Media**

We now investigate the relationship between influenza awareness on Twitter and the volume of influenza-related news media, using weekly counts from our newspaper dataset.

We first examined the correlation between awareness and news media volume, finding that they are highly correlated: across all ten regions, the mean Pearson correlation coefficient was .905 (SD .044), comparing each region's Twitter awareness value with each region's media count. Awareness and media volumes are even more correlated at the national level, at .94.[5] The weekly national media volumes can be seen alongside the weekly Twitter levels in Figure 1(b).

For comparison, the mean Pearson correlation between Twitter awareness and Twitter infection was .907 (SD .026), and .900 at the national level. This is about the same as me-

dia at the regional level, but weaker than the correlation with media at the national level.

Next, we investigated the relationship between awareness, infection, and media. Does awareness of flu track infection rates, or is it driven by media coverage? To answer this question, we used a bivariate linear regression model which estimates each week's Twitter awareness level as a linear combination of the week's infection level and news media volume, in a given region. Specifically, we use the following model:

$$awareness_{rw} = \beta_{r0} + \beta_{r1}\text{infection}_{rw} + \beta_{r2}\text{media}_{rw} + \epsilon_{rw} \quad (1)$$

where $\beta_{r0}$ is a region-specific intercept and each $\epsilon_{rw} \sim \mathcal{N}(0, \sigma_{rw}^2)$ is a residual term. The subscript $r$ denotes the region and $w$ denotes the week.

We fitted separate models for each region, as well as for national levels. Additionally, we experimented with two different definitions of infection$_{rw}$, fitting separate models for each: infection levels estimated from Twitter, using the same classifier used to measure awareness, as well as official influenza levels from the CDC's ILINet dataset. We standardized all data by replacing their values with z-scores, so that all values have the same unit (standard deviation).

Table 2 shows the learned model coefficients $\beta$. We see that news media volume has a higher coefficient than infection in all ten regions when using the ILINet infection variable. When using the Twitter infection variable, the infection coefficient has a higher contribution, though media still has a higher coefficient in five of ten regions, as well as at the national level. These results show that (i) as a function of infection and media volume, media generally contributes more to the level of public awareness, and (ii) compared to CDC ILINet infection levels, Twitter infection trends are closer to Twitter awareness trends.

**National versus regional media**   While the above model used news media counts from regional newspapers, we also compared to national media. Specifically, we used a similar regression model to estimate weekly regional awareness counts as a function of both the regional media volume and national media volume. After fitting the model, regional media had a mean coefficient of -.088 (SD .582) across the ten region-specific coefficients, while national media had a

---

[5]If we exclude tweets containing URLs, the correlation between awareness tweets and media drops to .888. This suggests that the high correlation is partly due to the sharing of links to news media, rather than more general awareness. This distinction between media sharing and other awareness is worthy of additional study in future work.

| | Twitter Infection | | CDC ILINet | |
|---|---|---|---|---|
| **Region** | **Infection** | **Media** | **Infection** | **Media** |
| 1 | .898 | .034 | .098 | .761 |
| 2 | .484 | .514 | .143 | .873 |
| 3 | .614 | .359 | .196 | .776 |
| 4 | .386 | .652 | .192 | .869 |
| 5 | .547 | .431 | .073 | .847 |
| 6 | .173 | .818 | -.003 | .978 |
| 7 | .341 | .645 | .119 | .852 |
| 8 | .580 | .401 | .161 | .785 |
| 9 | .490 | .531 | .186 | .834 |
| 10 | .561 | .435 | .228 | .741 |
| National | .340 | .645 | .0281 | .924 |

Table 2: Coefficients learned from two bivariate regression models that estimate each week's flu awareness level (as measured from Twitter) as a linear combination of the week's flu infection level and the week's level of media attention (as measured by newspaper volume). The first model uses the Twitter-based estimate of flu infection, while the second model uses the CDC's ILINet estimate.

mean coefficient of 1.026 (SD .561). Thus, national media explains regional awareness more than regional news media.

## Discussion and Conclusion

While prior work has established that Twitter is a valid source of real-time disease infection data, this study suggests that Twitter can also be used to understand disease *awareness*. We have analyzed trends in flu-related Twitter data separately for messages classified as infection versus awareness. We observed striking differences between these trends, and we conducted additional experiments to estimate the effect of infection and media on influenza awareness. Our key findings are:

- Infection tweets are substantially more correlated with CDC data than awareness tweets, as was the goal of the classifier. This highlights that the infection vs. awareness distinction is important for influenza surveillance.

- Influenza awareness is a function of news media volume more than infection levels. Moreover, media volume from national news outlets contribute much more to regional awareness levels than media from local newspapers within the regions.

- Likely for the reason that awareness is driven by national media, there is little regional variation in awareness trends.

- Awareness does not rise until the flu becomes severe, with awareness levels staying low even while infection levels rise. Additionally, awareness levels drop sharply after the peak, even when infection levels stay high. This might indicate that after the national highlighting is over, people quickly lose interest.

These observations only apply to the 2012–13 influenza season. Further analysis will be needed to determine if these conclusions can be generalized to other flu seasons. However, a possible conclusion from these findings is that there exists ample opportunity to drive preparation for epidemics by targeting only certain distribution channels, as national awareness strongly correlates with regional awareness.

Clearly, there is a complex relationship between flu prevalence, flu-related media attention, and public awareness of flu, and this relationship is not fully modeled by our simple regression analysis. Nevertheless, these experiments are an important step toward understanding these phenomena. Our findings show that awareness and infection are not as related as one might expect, and our experiments point to news media as an interesting confounder worthy of additional study.

## References

Aramaki, E.; Maskawa, S.; and Morita, M. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *EMNLP*.

Broniatowski, D. A.; Paul, M. J.; and Dredze, M. 2013. National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE* 8(12).

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *ACM Workshop on Soc.Med. Analytics*.

Dredze, M.; Paul, M. J.; Bergsma, S.; and Tran, H. 2013. Carmen: A Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Dredze, M.; Cheng, R.; Paul, M.; and Broniatowski, D. 2014. Healthtweets.org: A platform for public health surveillance using Twitter. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*.

Funk, S.; Gilad, E.; Watkins, C.; and Jansen, V. A. 2009. The spread of awareness and its impact on epidemic outbreaks. *Proc. Natl. Acad. Sci. U.S.A.* 106(16):6872–6877.

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.

Granell, C.; Gomez, S.; and Arenas, A. 2013. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Phys. Rev. Lett.* 111(12):128701.

Jones, J. H., and Salathe, M. 2009. Early assessment of anxiety and behavioral response to novel swine-origin influenza A(H1N1). *PLoS ONE* 4(12):e8032.

Lamb, A.; Paul, M. J.; and Dredze, M. 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter. *NAACL*.

Lampos, V., and Cristianini, N. 2012. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.* 3(4):1–22.

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The parable of Google Flu: Traps in big data analysis. *Science* 343(6167):1203–1205.

Paul, M. J.; Dredze, M.; and Broniatowski, D. 2014. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Preis, T., and Moat, H. S. 2014. Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science* 1(2).

Santillana, M.; Zhang, D. W.; Althouse, B. M.; and Ayers, J. W. 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am J Prev Med* 47(3):341–347.

Yuan, Q.; Nsoesie, E. O.; Lv, B.; Peng, G.; Chunara, R.; and Brownstein, J. S. 2013. Monitoring influenza epidemics in China with search query from Baidu. *PLoS ONE* 8(5).