Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns and their Impact on Surveillance Estimates

Jared Mowery

The MITRE Corporation

Abstract

Background: Influenza (flu) surveillance using Twitter data can potentially save lives and increase efficiency by providing governments and healthcare organizations with greater situational awareness. However, research is needed to determine the impact of Twitter users' misdiagnoses on surveillance estimates.

Objective: This study establishes the importance of Twitter users' misdiagnoses by showing that Twitter flu surveillance in the United States failed during the 2011-2012 flu season, estimates the extent of misdiagnoses, and tests several methods for reducing the adverse effects of misdiagnoses.

Methods: Metrics representing flu prevalence, seasonal misdiagnosis patterns, diagnosis uncertainty, flu symptoms, and noise were produced using Twitter data in conjunction with OpenSextant for geo-inferencing, and a maximum entropy classifier for identifying tweets related to illness. These metrics were tested for correlations with World Health Organization (WHO) positive specimen counts of flu from 2011 to 2014.

Results: Twitter flu surveillance erroneously indicated a typical flu season during 2011-2012, even though the flu season peaked three months late, and erroneously indicated plateaus of flu tweets before the 2012-2013 and 2013-2014 flu seasons. Enhancements based on estimates of misdiagnoses removed the erroneous plateaus and increased the Pearson correlation coefficients by .04 and .23, but failed to correct the 2011-2012 flu season estimate. A rough estimate indicates that approximately 40% of flu tweets reflected misdiagnoses.

Conclusions: Further research into factors affecting Twitter users' misdiagnoses, in conjunction with data from additional atypical flu seasons, is needed to enable Twitter flu surveillance systems to produce reliable estimates during atypical flu seasons.

Keywords: biosurveillance, social media, natural language processing, supervised machine learning

Correspondence: jmowery@mitre.org

DOI: 10.5210/ojphi.v8i3.7011

Introduction

Many studies have investigated using social media data or online data to perform biosurveillance [1, 2]. Eysenbach [3] was the first to use trends in internet searches as a means of estimating flu prevalence, and Ritterman et al. [4] subsequently became the first to use Twitter data for flu surveillance.

Twitter flu surveillance systems generally rely on keyword filters and classifiers to produce weekly counts of tweets indicative of flu prevalence. Lamb et al. [5] developed a classifier which distinguishes between tweets reflecting an awareness of the flu and tweets describing an infection with the flu, which tightens the causal relationship between weekly counts of flu tweets and Centers for Disease Control (CDC) or WHO measurements. Smith et al. [6] demonstrated that tweets related to general awareness of the flu yield substantially different trends than tweets related to infections, and Nagar et al. [7] reported that a classifier incorporating an annotator's estimate of the likelihood that a tweet indicated illness was important for their analysis of flu prevalence in New York City. Zuccon et al. [8] tested a wide variety of classifier types, with results indicating the choice of classifier has a limited effect on accuracy.

Recent studies have expanded the Twitter flu surveillance systems in a variety of ways, including encompassing multiple countries [9, 10], combining multiple indicators [10, 11], increasing geospatial resolution [7, 12–14], handling additional languages [15, 16], and estimating the secondary attack rate and serial interval [17].

However, Twitter flu surveillance relies on Twitter users' diagnoses of the flu. There are many potential causes of misdiagnoses. Nsoesie and Brownstein [1] observe that many existing systems likely measure influenza-like illness (ILI), which can be caused by a variety of non-flu pathogens. Chew and Eysenbach's Twitter content analysis during the 2009 pandemic [18] contains a rich set of metrics reflecting emotion levels, misinformation, and news or blog links that could all influence Twitter authors in choosing whether to tweet about an infection, and whether to diagnose that infection as the flu.

Since Twitter is not a representative sample of the United States' population [19-21], Twitter flu surveillance estimates will be biased. Studies have investigated potential variations in the peak time, morbidity, and rate of flu transmission as a function of age

group and social networks [22-25]. Region and humidity may also influence flu mortality rates and spread [26-27]. Finally, although positive specimen counts for the CDC or WHO are used as ground truth data, variations in the collection and testing of specimens, participation levels of laboratories, and other factors may introduce sampling biases.

Detecting atypical flu seasons reliably is important, since they may require atypical responses from governments and healthcare organizations to save lives and increase efficiency. This study focuses on flu seasons with atypical onset times, such as the 2011-2012 flu season, since these yield the most direct evidence for misdiagnoses. Since this study is intended to quantify Twitter users' misdiagnoses rather than maximize the correlation between flu estimates and WHO counts, it does not incorporate additional data sources which could obscure misdiagnosis patterns in Twitter, such as search query volumes or time-lagged positive specimen count data. Many of the algorithms were implemented using the R Project for Statistical Computing [28].

Methods

Data Collection and Classification

This study used Gnip Decahose [29] data, which is a 10% pseudo-random sample of publicly available tweets. The tweet volumes collected each week between the weeks starting on 2011-08-01 and 2014-09-15 exhibit several gaps due to internet connectivity issues and hardware failures. These gaps were corrected by extrapolating from nearby data using a two pass process.

The first pass applied a sliding median filter of width 15 to approximate the expected counts for each week. Any range of weeks with week indices [a, b] in which zero tweets were collected was replaced by the estimated values from a linear interpolation between the values at indices $a - 2$ and $b + 2$.

The second pass applied a sliding median filter of width 7 to the results of the first pass. The following equation was used to produce a corrected count $\hat{t}_i$ for each week i:

$$\hat{t}_i = \begin{cases} s_i \text{ if } t_i < 0.9 s_i, \\ t_i \text{ otherwise.} \end{cases} \quad (1)$$

where s_i is the output of the second sliding median filter and t_i is the tweet count after zeroes were replaced by the first pass. The constant 0.9 was chosen to apply the correction only when the weekly count was at least 10% less than the expected count, which served as a rough method for identifying weeks during which data loss occurred. Applying Equation 1 compensated for the gaps in data collection (Figure 1).

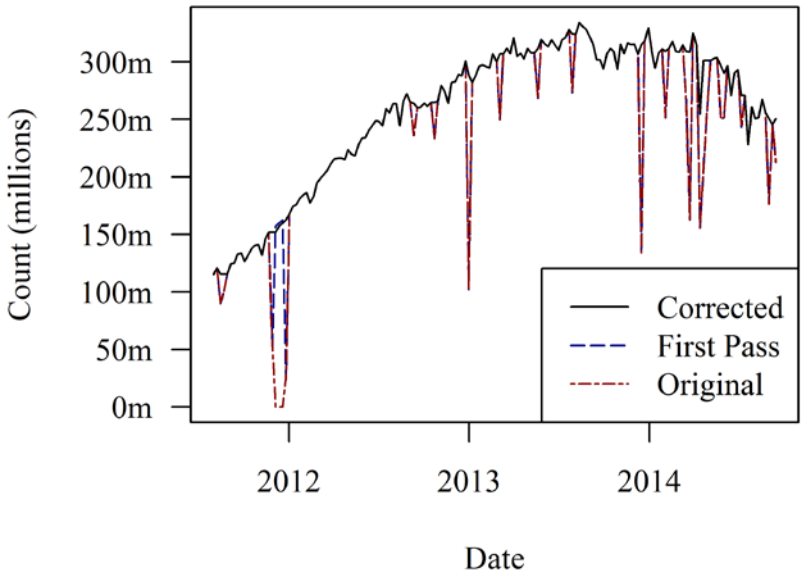

Figure 1: Tweets collected per week. The Original series shows the number of tweets collected from the Decahose feed. The First Pass series depicts the result of using linear interpolation to replace the counts for weeks in which zero tweets were collected. The Corrected series shows the estimated number of tweets which would have been collected for each week if there had not been data collection gaps.

The metrics based on Twitter data must also be adjusted to compensate for the data losses. The following equation produced adjusted counts for each week *i*:

$$\hat{c}_i = \begin{cases} \hat{c}_k \text{ if } (t_i \text{ or } t_{i-1} \text{ or } t_{i+1} = 0), \text{with the maximum } k \text{ s.t. } k < i \text{ and } t_k > 0, \\ c_i \dfrac{\hat{t}_i}{t_i} \text{ if } t_i \neq \hat{t}_i \text{and } t_i > 0, \\ c_i \text{ otherwise.} \end{cases} \quad (2)$$

where c_i is the count produced by a metric and $\hat{c}_i$ is the count adjusted for potential data loss. This equation assumes the fraction of tweets which match the criteria for a metric is consistent, so the value of the metric during a week which experienced data loss can be approximated by applying the same fraction to the number of tweets expected during that week. For weeks in which no tweets were collected, the adjusted metric value for the most recent week in which tweets were collected was used. Although a better estimate could have been obtained through linear interpolation, this approach uses only data which would have been available at the time.

This study used the WHO's weekly positive counts of flu virus specimens in the United States, including types A and B [30], as ground truth data. The 2011-2012 flu season peaked approximately three months late compared to the 2012-2013 and 2013-2014 flu seasons. This is valuable for quantifying the extent to which Twitter users' misdiagnoses adversely affect the correlation strength between Twitter flu surveillance estimates and WHO positive specimen counts, since tweets in late 2011 most likely reflect misdiagnoses.

The maximum entropy classifier was trained on 1,274 English language tweets containing illness or symptom related terms collected between December 31, 2011 and January 31, 2012. Each tweet was hand-annotated by a single annotator for indications that the author, or someone the author knew, was ill. Examples of illness included flu, common colds, allergies, and symptoms such as nausea, sore throat, and nasal congestion. Instances of symptoms not due to illness, such as nausea due to overeating, stomach pain due to consuming spicy foods, and muscle aches due to exercise, were not counted as illness. The tweets which were related to illness according to the classifier are referred to as "sick tweets" in this paper. Due to the expense of developing classifiers for multiple languages, non-English tweets were not considered in this study.

The maximum entropy classifier used Apache's OpenNLP [31] implementation. Retweets and tweets containing URLs were excluded to help reduce the number of tweets related to news stories or memes. Unigrams, bigrams, and the tweet length in [0.0, 1.0], with 1.0 corresponding to a length of 200 characters, were used as features since they are commonly used and computationally inexpensive. The classifier used Gaussian regularization with $\sigma = 1.0$ and 10,000 iterations to ensure convergence. The classifier's performance was tested using stratified 10-fold cross-validation. To bias the classifier in favor of precision over recall, only tweets whose classifier score exceeded 0.75 were designated as sick tweets. The constant 0.75 was chosen since it yielded weekly counts typically over 100 for sick tweets which contained the word "flu". The lowest non-zero weekly count was 97, and the average count was 696.

Metrics Collection

This study collected several metrics from the sick tweets. Tweets were filtered using illness and symptom related keywords, restricted to the United States by applying OpenSextant [32] to the user-provided location fields, and then limited to the English language using the Cybozu Labs Language Detection Library for Java [33]. Out of the 13,273,284 tweets containing illness or symptom related terms, OpenSextant provided estimated locations for 3,667,309 of them, or 27.6%. Retweets and tweets containing URLs were excluded to match the classifier training data.

Table 1: Case-insensitive queries used to define each metric. Each metric is restricted to English tweets classified as sick tweets from the United States.

	Query	Example
Flu	Flu	Feeling miserable. Go away flu!
Uncertainty	might or maybe or hope	I might be coming down with a fever
UncertaintyF	(might or maybe or hope) and flu	Sore throat… nose like a tap… might be flu
Symptom	sore throat or fever	Had a sore throat for days now
SymptomF	(sore throat or fever) and flu	Fever all day, hope it's not flu

Most of the metrics were simply defined as the fraction of tweets each week which matched a case insensitive query (Table 1). The Flu metric contained only sick tweets with the word "flu", which are referred to as "flu tweets" in this paper. The Uncertainty metric is intended to measure Twitter authors' uncertainty in their diagnoses, such as "I might be getting sick", "Maybe this is just an allergy", or "I hope this is not the flu". The Symptom metric measures tweets containing two common symptoms of influenza-like illness: fevers and sore throats. Finally, metrics with the suffix "F" have been restricted to flu tweets. Since the weekly counts of flu tweets were generally over 100, this study did not examine misspellings of query terms or the use of slang.

The Noise metric is an estimate of the expected fraction of flu tweets during periods in which the flu is not prevalent. The thirteen weeks occurring in the middle of each year were used to estimate the noise level, which corresponds to an estimate that approximately one quarter of weeks during the year are not substantially affected by the flu season. The mean count for each of these midyear periods was used as a noise estimate. Due to the difficulty of distinguishing flu tweets arising from flu infections from tweets arising from misdiagnoses, noise cannot effectively be measured during periods in which the flu is prevalent. Therefore, each consecutive pair of midyear noise estimates was linearly interpolated to generate the complete noise estimate. The noise level gradually decreased during the period tweets were collected, which may be a consequence of the atypical 2011-2012 flu season (Figure 2).

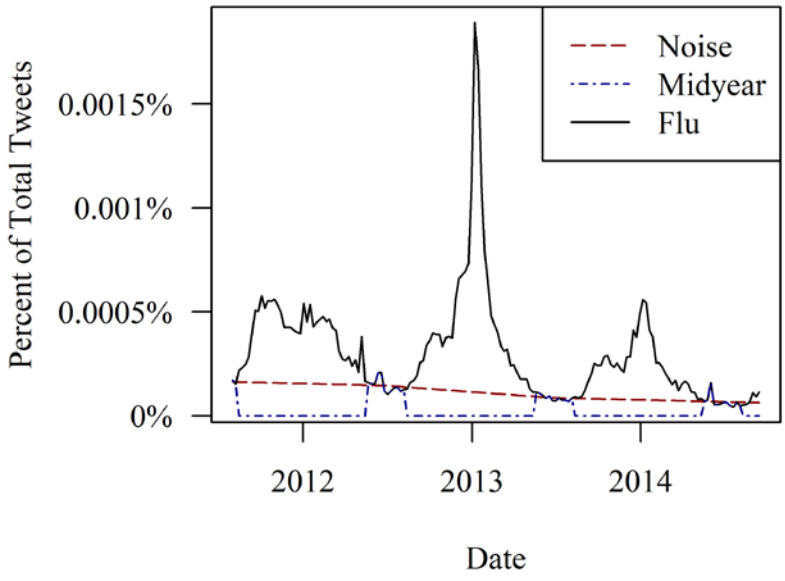

Figure 2: Noise estimate based on linearly interpolating noise estimates from each midyear period. The Midyear series shows the weeks which were used to estimate the noise for each midyear period. Each series has been divided by the corrected total number of tweets collected each week.

Misdiagnosis Measurement

Since WHO positive specimen counts show the flu was not prevalent from August 2011 through December 2011, despite an increase in flu tweets, the flu tweets from that time period largely represent misdiagnoses. Measuring the number of misdiagnosis tweets over time for a typical flu season is potentially valuable for counteracting their effects on Twitter flu surveillance, but there are two major challenges:

1) separating the misdiagnosis tweets from the small number of correct diagnoses of the flu, classifier false positives, and other sources of noise from August 2011 to December 2011, and
2) estimating misdiagnosis tweets for January 2012 through May 2012, since direct measurement is complicated by the genuine prevalence of the flu.

To address the first challenge, this study subtracts the Noise metric from the Flu metric. The Noise metric is an estimate of the fraction of flu tweets expected during periods in which the flu is not prevalent. Since the flu was not prevalent in late 2011, the Flu metric should have equaled the Noise Metric during that time period. Therefore, subtracting the Noise metric leaves the flu tweets which contributed to the unexpected rise in flu tweets during late 2011.

To address the second challenge, this study estimates misdiagnosis tweets from late 2011 and extrapolates them to early 2012. The weekly fractions of misdiagnosis tweets from August to December 2011 were estimated by smoothing the flu tweets, subtracting the Noise metric, and normalizing by the Noise metric:

$$m_i = \frac{med(f_i) - n_i}{n_i} \quad (3)$$

where i is the week (limited to August through December 2011), m_i is a unitless factor which estimates the fraction of misdiagnosis tweets when multiplied by the Noise metric, *med* is a sliding median filter of width 5, f_i is the flu metric, and n_i is the Noise metric. Both f_i and n_i are expressed as fractions of the corrected total tweet count for week i. The smoothing is intended to reduce the effects of noise, and the normalization by n_i helps account for factors which may change from season to season by assuming the misdiagnosis estimate is proportional to the noise estimate.

This study hypothesized two extrapolations based on m: Tapered and Symmetric. The Tapered extrapolation assumes misdiagnosis tweets taper off as the flu season progresses, which continues the downward trend seen in misdiagnosis tweets at the end of 2011. The tapering was implemented with a linear interpolation between the misdiagnosis fraction at the end of 2011 (week starting 2012-01-02) and the estimate of the noise baseline at the end of the flu season (week starting 2012-06-04). Tapering could be caused by psychosocial factors, such as decreasing anxiety due to news media coverage reporting that the flu season was mild or late. The Symmetric extrapolation assumes the misdiagnosis tweet pattern is symmetric around the end of 2011, and the symmetry was implemented by concatenating the weekly counts in the weeks [2011-08-01, 2012-01-02] with the reversed weekly counts in weeks [2011-08-01, 2011-12-26]. The symmetric extrapolation assumes misdiagnosis tweets do not taper off as the flu season progresses, and that Twitter authors' misdiagnoses are symmetric around the typical peak of a flu season. This could correspond to Twitter users' misdiagnoses reflecting their expectations of flu prevalence during a typical flu season. Both estimates of the misdiagnosis errors cover the same range of weeks.

Copying the unitless estimates m_i and the extrapolated values (weeks 2011-08-01 to 2012-06-04) to the corresponding weeks centered on January 1st of the 2012-2013 (weeks 2012-07-30 to 2013-06-03) and 2013-2014 (weeks 2013-07-29 to 2014-06-02) flu seasons, and then multiplying by the Noise metric, yielded the final estimate of the fraction of misdiagnosis tweets for 2011-2014 (Figure 3). Since the misdiagnosis estimate was constructed to be proportional to the noise estimates from the midyear periods, and since those midyear periods were likely to have few tweets correctly diagnosing the flu, the midyear periods were excluded from the misdiagnosis estimates.

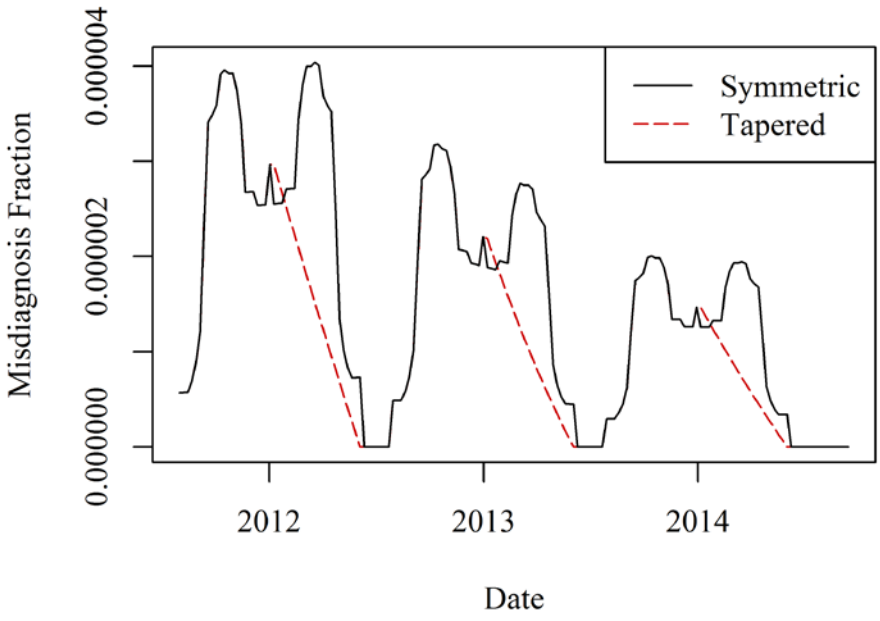

Figure 3: Estimated weekly fraction of misdiagnosis tweets.

Finally, the two misdiagnosis based estimates of flu prevalence were produced by subtracting the weekly estimates of the fraction of misdiagnosis tweets from the weekly fraction of flu tweets for each of the two extrapolations.

Misdiagnosis Cross-Validation

The previous section used the prior knowledge that WHO positive specimen counts for late 2011 are approximately equal to the positive specimen counts when flu is not prevalent. However, this means its results can only be tested against data from early 2012 onward, or that it must rely on comparisons with recent WHO positive specimen counts. Therefore, this study also uses a form of 3-fold cross-validation, in which an estimate is produced for a "test" flu season by using misdiagnosis tweet rates estimated by taking the difference between the WHO positive specimen counts and fractions of flu tweets for the remaining two "training" flu seasons. For each flu season, the same range of weeks was used as in the previous section.

However, this approach requires comparing positive specimen counts and fractions of flu tweets. This paper used a simple linear regression, $P \sim cF$, between the WHO positive specimen counts (P) and the fraction of flu tweets for the non-test weeks (F) to obtain a constant (c) representing a best estimate of the unit conversion factor. The linear regression did not include a constant term, so the linear regression only estimated the single coefficient *c*.

$$m_i = \left(\frac{cf_i - p_i}{c} - n_i\right) \times {n_i}^{-1} \quad (4)$$

Equation 4 details obtaining the unitless misdiagnosis estimate m_i for a flu season, where i is the week, c is the coefficient for unit conversion obtained via linear regression, f_i is the Flu metric, p_i is the positive specimen count from the WHO, and n_i is the Noise metric. The final misdiagnosis tweet fraction estimate for the test flu season was obtained by averaging the unitless misdiagnosis estimates for the two training flu seasons and multiplying by the Noise metric for the test flu season. The misdiagnosis tweet fraction estimate was subtracted from the test flu season's weekly fractions of flu tweets to yield the final estimate of flu prevalence.

Results

Data Collection and Classification

The maximum entropy classifier achieved an F-measure of .76, with .73 precision and .79 recall. There were 354 true positives compared to 129 false positives, and 697 true negatives compared to 94 false negatives. To produce the actual counts of sick tweets, the classifier's threshold was increased to .75 to favor precision over recall, since precision is more important for this study. The .75 threshold achieved an F-Measure of .72, with .86 precision and 0.61 recall.

The Pearson correlation coefficient between the sick tweets and the WHO's positive specimen counts is r = .66 (P < .001), which demonstrates that there is a significant degree of correlation even before filtering the sick tweets to examine only flu tweets.

Metrics

The Flu metric achieved a Pearson correlation with the WHO positive specimen counts of r = .72 (P < .001), which is an improvement over the correlation for sick tweets of r = .66. However, the Flu metric erroneously reports a typical flu season occurring in late 2011 and early 2012, as well as plateaus of flu tweets occurring prior to the start of the next two flu seasons (Figure 4). The 2011-2012 flu season is erroneous in the sense that there is a substantial rise in flu tweets in late 2011 despite the lack of a corresponding increase in WHO positive specimen counts, resulting in the flu tweets exhibiting a pattern of elevated counts roughly centered on December even though the actual flu season peak occurred months later, according to the WHO positive specimen counts.

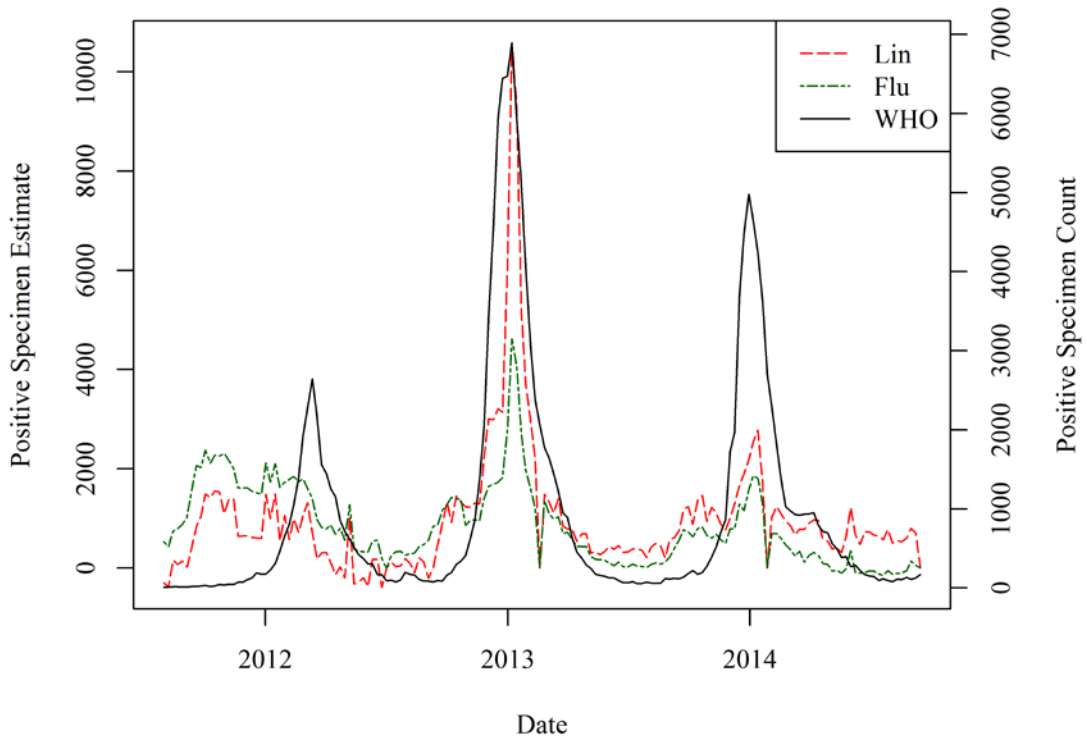

Figure 4: Flu prevalence estimates versus WHO positive specimen count data (WHO) for the linear combination of the flu, noise, and uncertain metrics (Lin), and the flu metric alone (Flu). Although the Uncertain metric improves the correlation, both the flu and linear combination results erroneously estimated a 2011-2012 flu season occurring at the typical time, and produced plateaus of misdiagnosis tweets before each subsequent flu season.

To measure the relative efficacy of the remaining metrics, the Pearson correlation coefficients between linear regressions of the metrics and the WHO positive specimen count data were calculated (Table 2). In each case, the linear regression included a constant term. To reduce over-fitting, each calculation used 10-fold cross-validation, in which the folds were obtained by partitioning the date range into 10 approximately equal-length time periods. The combination of using 10-fold cross-validation and linear regression increased the difficulty of obtaining high correlation coefficients, which reduced the correlation for the Flu metric from $r = .72$ to $r = .54$. Introducing the Noise metric substantially improved the correlation result, while adding the Sick tweets metric yielded no additional benefit. Holding the number of regressors constant by substituting the other metrics for the Sick metric revealed that only the Uncertain metric provided a substantial benefit.

Table 2: Pearson correlation coefficients for multiple variable linear regressions using 10-fold cross-validation. The Uncertain metric substantially increases the correlation with the WHO's positive specimen count. Note: the correlation for the Flu metric is 0.72 when not using 10-fold cross-validation and multiple variable linear regression.

	R
Flu	.54
Flu + Noise	.73
Flu + Sick + Noise	.73
Flu + Uncertain + Noise	.77
Flu + UncertainF + Noise	.73
Flu + Symptom + Noise	.72
Flu + SymptomF + Noise	.72

While the Uncertain metric improved the correlation coefficient, the regressions failed to remove the misdiagnosis tweets, which erroneously indicated a typical 2011-2012 flu season and erroneously showed plateaus of flu activity occurring before each of the next two flu seasons (Figure 4).

Misdiagnosis Measurement

The Flu, Symmetric, and Tapering metrics all correlate with the WHO's ILI positive specimen counts (Table 3). The sum of *P* values for each correlation in the table was $P < .001$, indicating that the set of correlations passes the Bonferroni correction. However, the metrics vary in correlation strength: the Flu metric suffers from significant plateaus of misdiagnosis tweets preceding each flu season, the Symmetric metric can be rejected since it produces flu estimates below the noise baseline during each of the three flu seasons, and the Tapering metric successfully removes the false positive plateaus preceding each flu season but shows the flu seasons starting late (Figure 5). The Tapering metric achieved slightly higher correlations than the other two metrics in all three test conditions, and the Tapering metric gains the most benefit when more of the atypical 2011-2012 flu season is included in the test. However, the test which excludes none of the data from the 2011-2012 season is only included for reference; since the late 2011 tweets were used to construct the misdiagnosis tweets estimate, using that data comingles tuning and testing data.

Table 3: Pearson correlation coefficients for the flu metric as well as the flu metric after subtracting the Symmetric and Tapering estimates of misdiagnosis tweets. The rows present the correlations when excluding none of the data, the first half of the typical 2011-2012 flu season, or the entire 2011-2012 flu season. Flu tweets from late 2011 were used to measure the misdiagnosis tweets, and are included in the row for excluding none of the data.

Exclusion	Flu	Symmetric	Tapering
None	.72	.73	.81
Half	.82	.77	.83
2011-2012	.84	.83	.85

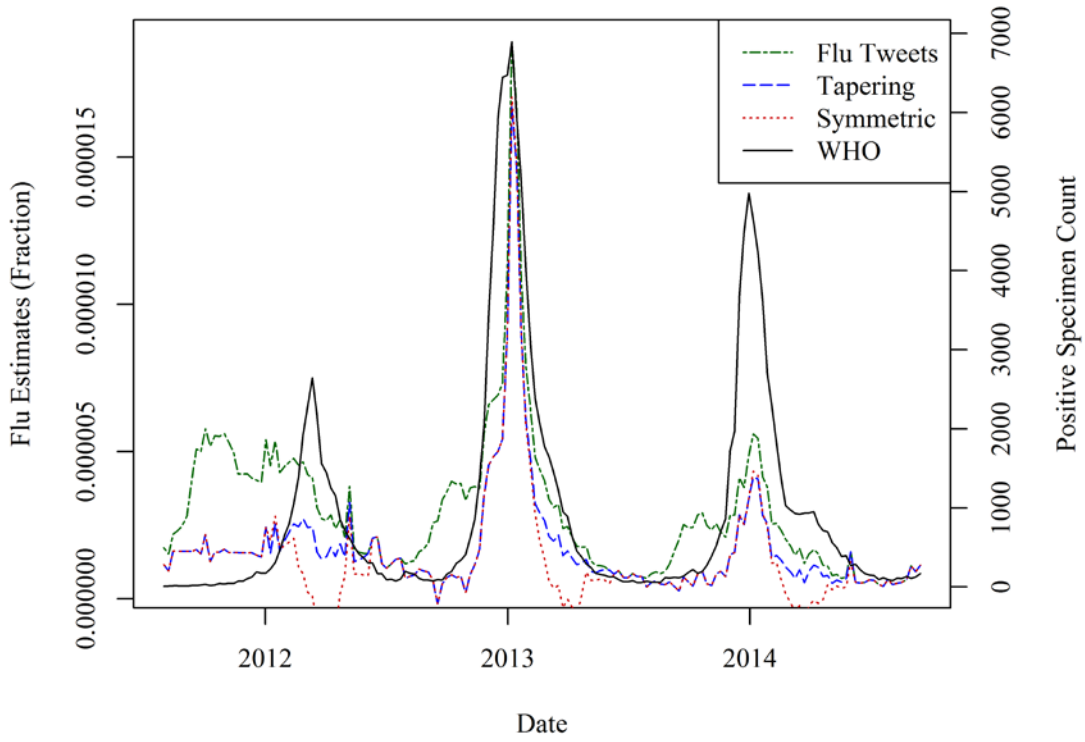

Figure 5: Estimated flu prevalence before and after subtracting estimated misdiagnosis tweets for each of the Tapering and Symmetric extrapolation methods. The Symmetric method can be rejected since it produces flu estimates below the noise level for all three flu seasons. The Tapering method successfully removes the plateaus of misdiagnosis tweets which precede each of the three flu seasons, but shows the 2012-2013 and 2013-2014 flu seasons starting late. The Tapering and Symmetric methods frequently overlap in the plot, due to sharing the same weekly misdiagnosis estimates for late 2011.

The Tapering metric indicates that approximately 47,907 tweets were misdiagnoses, although this may be an overestimate since the 2012-2013 and 2013-2014 flu seasons

start late according to the Tapering metric. There were 121,234 flu tweets total, which suggests that roughly 39.52% of the flu tweets reflected misdiagnoses.

Misdiagnosis Cross-Validation

Removing estimated misdiagnosis tweets based on 3-fold cross-validation for the three flu seasons successfully removes the plateaus of misdiagnosis tweets occurring before the 2012-2013 and 2013-2014 flu seasons, while accurately reflecting the correct start dates for the 2012-2013 and 2013-2014 flu seasons (Figure 6). However, the erroneous estimate for the 2011-2012 flu season remains. The Pearson correlation coefficient was $r = .76$ ($P < .001$), compared to $r = .72$ for the Flu metric.

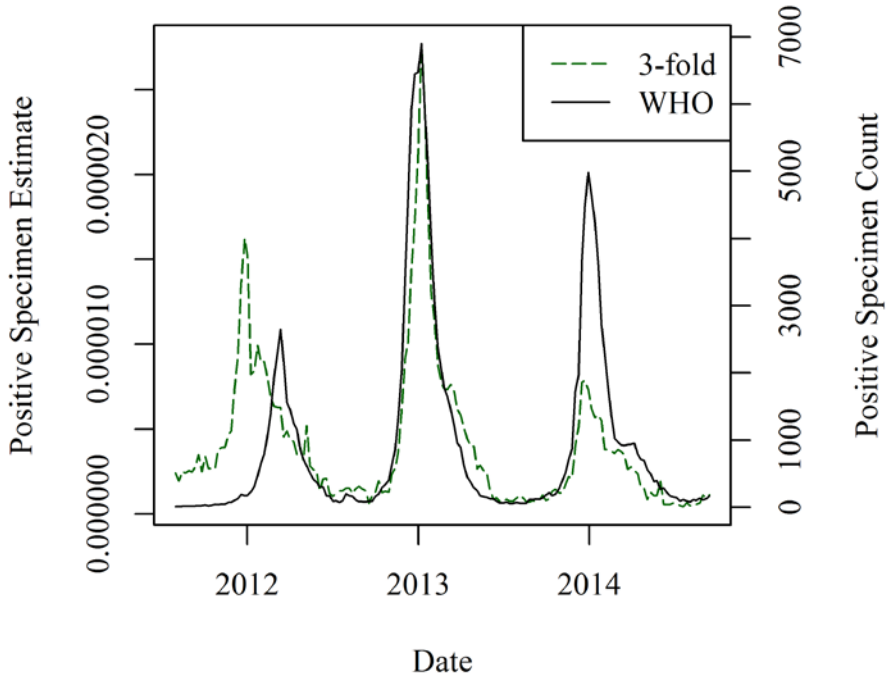

Figure 6: Comparison of the Flu metric, after subtracting the 3-fold misdiagnosis estimate, to WHO positive specimen counts. The 3-fold estimate successfully removes the plateaus of flu tweets occurring prior to the starts of the 2012-2013 and 2013-2014 flu seasons, and accurately reflects the start dates of the 2012-2013 and 2013-2014 flu seasons, but it is unable to remove sufficient misdiagnosis tweets from the 2011-2012 flu season to reveal the season's atypical timing.

Discussion

This study establishes the importance of misdiagnoses by showing that the pattern of flu tweets during the 2011-2012 flu season fails to approximate the WHO positive specimen counts, and that the flu tweets exhibit plateaus of misdiagnosis tweets preceding each of the next two flu seasons. This study quantifies the importance of misdiagnosis tweets by showing that the Tapering metric increases the correlation coefficient from $r = .72$ for the

flu metric alone to r = .81, removes the plateaus of misdiagnosis tweets prior to the 2012-2013 and 2013-2014 flu seasons, and yields an estimate that 39.52% of flu tweets (47,907 / 121,234) reflect misdiagnoses. Finally, this study demonstrates that misdiagnoses can be counteracted via the Uncertain and Noise metrics (r = .54 increased to r = .77) and by applying 3-fold cross-validation to produce an estimate of seasonal misdiagnosis patterns (r = .76).

However, each approach has limitations. Only the Tapering metric enabled detection of the 2011-2012 flu season, and it was developed with the prior knowledge that WHO positive specimen counts in late 2011 were low. This is useful for quantifying the impact of misdiagnoses, but presents a challenge for non-retrospective flu surveillance. While an implementation could use time-lagged WHO counts and apply the Tapering metric only once the flu season began, this may not be robust and it would sacrifice the ability to detect the start of the flu season via Twitter data. Non-retrospective flu surveillance can be enhanced by using either the Uncertain and Noise metrics or the 3-fold cross-validation estimate of seasonal misdiagnosis patterns. However, only the latter successfully removed misdiagnosis tweet plateaus before the 2012-2013 and 2013-2014 flu seasons, which is necessary to accurately detect the beginnings of the 2012-2013 and 2013-2014 flu seasons.

The limited availability of Twitter data in atypical flu seasons is a significant challenge for further analysis of misdiagnosis tweets. Analyzing multiple countries during an atypical flu season may be beneficial, but evidence that flu is spread by air travel [34] means that results for each country could not be treated as statistically independent.

Further research could address improvements to data collection and classification, such as developing classifiers for multiple languages, experimenting with more complex classifiers and feature extraction, examining the effects of different annotation guidelines, using larger volumes of annotated tweets, and using expanded queries including misspellings and references to taking medications. In addition, demographic differences between Twitter users and WHO sampling may introduce additional inaccuracies. Finally, the data losses experienced during certain weeks of data collection may have produced inaccurate estimates despite the corrections described in the Methods section.

This study focused on quantifying seasonal misdiagnosis errors specifically in Twitter data, rather than incorporating multiple exogenous data sources or statistical techniques to obtain the best possible estimate of flu prevalence. Many studies have shown that using multiple data sources and applying a variety of models can improve flu estimates. As a recent example, Santillana et al. demonstrated that using a combination of time-lagged CDC data and a new, timely source of electronic health records, which are not available to the public, can improve the accuracy of flu surveillance systems [35].

Twitter flu surveillance research is promising, but identifying misdiagnosis tweets remains a challenge. Although this paper presents methods of enhancing Twitter flu surveillance for flu seasons by using estimates of seasonal misdiagnosis tweeting patterns, these same seasonal misdiagnosis patterns also indicate a risk that there is only a weak causal connection between individuals infected with the flu and Twitter authors reporting

flu infections. The weak causal connection is illustrated by the lack of correlation between flu tweets and WHO positive specimen counts during the 2011-2012 flu season, even after applying corrections for seasonal misdiagnosis patterns. Further research, in conjunction with data from additional atypical flu seasons, is needed to enable Twitter flu surveillance systems to produce reliable estimates of flu, rather than ILI, during atypical flu seasons.

Acknowledgements

The author would like to thank The MITRE Corporation for funding this research.

Conflicts of Interest

None declared. As a not-for-profit operator of federally funded research and development centers, The MITRE Corporation is not permitted to compete with industry.

References

1. Nsoesie EO, Brownstein JS. 2015. Computational approaches to influenza surveillance: beyond timeliness. *Cell Host Microbe*. 17(3), 275-78. PubMed http://dx.doi.org/10.1016/j.chom.2015.02.004

2. Paul MJ, Sarker A, Brownstein JS, Nikfarjam A, Scotch M, et al. Social Media Mining for Public Health Monitoring and Surveillance. Pacific Symposium on Biocomputing (PSB); 2016; Kohala Coast, Hawaii. 2016. pp. 468-79.

3. Eysenbach G. 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance in AMIA. *AMIA Annu Symp Proc*. •••, 244-48. PubMed

4. Ritterman J, Osborne M, Klein E. Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic. Proceedings of the 1st International Workshop on Mining Social Media; 2009; Seville, Spain. 2009. pp. 9-17.

5. Lamb A, Paul MJ, Dredze M. Separating Fact from Fear: Tracking Flu Infections on Twitter. HLT-NAACL; 2013; Atlanta, Georgia, USA. 2013. pp. 789–795.

6. Smith MC, Broniatowski DA, Paul MJ, Dredze M. Towards Real-Time Measurement of Public Epidemic Awareness: Monitoring Influenza Awareness Through Twitter. AAAI Spring Symposium on Observational Studies Through Social Media and Other Human-Generated Content; 2016; Stanford, California. 2016.

7. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, et al. 2014. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res*. 16(10), e236. PubMed http://dx.doi.org/10.2196/jmir.3416

8. Zuccon G, Khanna S, Nguyen A, Boyle J, Hamlet M, Cameron M. Automatic detection of tweets reporting cases of influenza like illnesses in Australia. Health Inf

Sci Syst 2015 Feb 24;3(Suppl 1 HISA Big Data in Biomedicine and Healthcare 2013 Con):S4. PMID:25870759

9. Paul M, Dredze M, Broniatowski D, Generous N. Worldwide Influenza Surveillance Through Twitter. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence; AAAI Conference on Artificial Intelligence; 2015; Austin, Texas. 2015.

10. Zhang Q, Gioannini C, Paolotti D, Perra N, Perrotta D, et al. Social Data Mining and Seasonal Influenza Forecasts: The FluOutlook Platform. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases; 2015; Porto, Portugal. 2015. pp. 237-240.

11. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, et al. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLOS Comput Biol*. 11(10), e1004513. PubMed http://dx.doi.org/10.1371/journal.pcbi.1004513

12. Dredze M, Paul MJ, Bergsma S, Tran H. Carmen: A Twitter Geolocation System with Applications to Public Health. Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence; AAAI Conference on Artificial Intelligence; 2013; Bellevue, Washington. 2013.

13. Broniatowski DA, Dredze M, Paul MJ, Dugas A. 2015. Using social media to perform local influenza surveillance in an inner-city hospital: a retrospective observational study. *JMIR Public Health Surveill*. 1(1), e5. PubMed

14. Broniatowski DA, Paul MJ, Dredze M. 2013. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS One*. 8(12), e83672. PubMed http://dx.doi.org/10.1371/journal.pone.0083672

15. Li J, Huang W, Chen P. LDA Based Event Extraction: Detecting Influenza Epidemics Using Microblog. The Second International Conference on Data Science; 2015; Sydney, Australia. 2015. pp. 30-33.

16. Sun X, Ye J, Ren F. Hybrid Model Based Influenza Detection with Sentiment Analysis from Social Networks. Proceedings of the 4th National Conference in Social Media Processing; 2015; Guangzhou, China. 2015. pp. 51–62.

17. Yom-Tov E, Johansson-Cox I, Lampos V, Hayward AC. 2015. Estimating the secondary attack rate and serial interval of influenza-like illnesses using social media. *Influenza Other Respi Viruses*. 9(4), 191-99. PubMed http://dx.doi.org/10.1111/irv.12321

18. Chew C, Eysenbach G. 2010. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One*. 5(11), e14118. PubMed http://dx.doi.org/10.1371/journal.pone.0014118

19. Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN. Understanding the demographics of Twitter users. ICWSM; 2011; Barcelona, Spain. 2011. pp. 554-557.

20. Hecht B, Stephens M. A tale of cities: Urban biases in volunteered geographic information. ICWSM; 2014; Ann Arbor, Michigan. 2014. pp. 197-205.

21. Malik M, Lamba H, Nakos C, Pfeffer J. Population bias in geotagged tweets. ICWSM; 2015; Oxford, England. 2015. pp. 18-27.

22. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, et al. 2015. Age-related differences in the accuracy of web query-based predictions of influenzalike illness. *PLoS One*. 10(5), e0127754. PubMed http://dx.doi.org/10.1371/journal.pone.0127754

23. Schanzer D, Vachon J, Pelletier L. 2011. Age-specific differences in influenza epidemic curves: do children drive the spread of influenza epidemics? *Am J Epidemiol*. 174(1), 109-17. PubMed http://dx.doi.org/10.1093/aje/kwr037

24. Glass LM, Glass RJ. 2008. Social contact networks for the spread of pandemic influenza in children and teenagers. *BMC Public Health*. 8, 61. PubMed http://dx.doi.org/10.1186/1471-2458-8-61

25. Nsoesie EO, Marathe M, Brownstein JS. 2013. Forecasting peaks of seasonal influenza epidemics. *PLoS Curr*. •••, 5. PubMed

26. Shaman J, Pitzer VE, Viboud C, Grenfell BT, Lipsitch M. 2010. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol*. 8(2), e1000316. PubMed http://dx.doi.org/10.1371/journal.pbio.1000316

27. Yang W, Lipsitch M, Shaman J. 2015. Inference of seasonal and pandemic influenza transmission dynamics. *Proc Natl Acad Sci USA*. 112(9), 2723-28. PubMed http://dx.doi.org/10.1073/pnas.1415012112

28. R Core Team. 2016. A Language and Environment for Statistical Computing R Foundation for Statistical Computing. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf. Archived at: http://www.webcitation.org/6iCHaoYyS

29. Gnip Inc. 2016. Decahose: Real Time Trend Detection and Discovery. https://gnip.com/realtime/decahose/. Archived at: http://www.webcitation.org/6jaw24k6R

30. World Health Organization. 2016. FluNet. http://www.who.int/influenza/gisrs_laboratory/flunet/en/. Archived at: http://www.webcitation.org/6jawE3GNL

31. Apache Software Foundation. 2016. The Apache OpenNLP Library. https://opennlp.apache.org/. Archived at: http://www.webcitation.org/6hvXhTr5U

32. Ubaldino M, Lutz D. 2015. Open Source Entity Extraction, Geocoding and Temporal Coding Tools. https://github.com/OpenSextant. Archived at: http://www.webcitation.org/6jawHinp2

33. Nakatani S. 2010. Language Detection Library for Java. https://github.com/shuyo/language-detection. Archived at: http://www.webcitation.org/6jawJ5DMA

34. Leitmeyer K, Adlhoch C. 2016. Influenza transmission on aircraft: a systematic literature review. *Epidemiology*. 27(5), 743-51. PubMed http://dx.doi.org/10.1097/EDE.0000000000000438

35. Santillana M, Nguyen AT, Louie T, Zink A, Gray J, et al. 2016. Cloud-based electronic health records for real-time, region-specific influenza surveillance. *Sci Rep*. 6, 25732. PubMed http://dx.doi.org/10.1038/srep25732