Examining the Causal Relationships Among Pollen Count, Tweet Numbers, and Patient Numbers for Seasonal Allergic Rhinitis Surveillance: A Retrospective Analysis of Multi-Source Data

Abstract

**Background:** Health-related social media data are increasingly used in disease surveillance studies, which have demonstrated moderately high correlations between the number of social media posts and the number of patients. However, there is a need to understand the causal relationship between the behavior of social media users and the actual number of patients to increase the credibility of disease surveillance based on social media data.

**Objective:** To clarify the causal relationships among pollen count, the posting behavior of social media users, and the number of SAR patients in the real world.

**Methods:** This analysis was conducted using datasets of pollen count, tweet numbers, and SAR patient numbers from Kanagawa Prefecture, Japan. We examined daily pollen counts for Japanese cedar (the major cause of SAR in Japan) and hinoki cypress (which commonly complicates SAR) from February 1 to May 31, 2017. The daily numbers of tweets that included the keyword "kafunshō" (or SAR) were calculated between January 1 and May 31, 2017. Daily SAR patient numbers from January 1 to May 31, 2017 were obtained from three healthcare institutes that participated in the study. The Granger causality test was used to examine the causal relationships among pollen count, tweet numbers, and the number of SAR patients during the study period from February to May 2017. To determine if time-variant factors affect these causal relationships, we also analyzed the main SAR phase (February to April) when Japanese cedar trees actively produce and release pollen.

**Results:** Increases in pollen count were found to cause increases in the number of tweets during the overall study period (P = .04), but not the main SAR phase (P = .05). In contrast, increases in pollen count were found to increase patient numbers in both the study period (P = .04) and the main SAR phase (P = .01). In addition,
increases in the number of tweets also caused increases in patient numbers during the main SAR phase (P = .02), but not the overall study period (P = .89). Patient numbers did not affect the number of tweets in both the overall study period (P = .24) and the main SAR phase (P = .47).

**Conclusions:** Understanding the causal relationships between these factors is an important step to increase the credibility of surveillance systems that use social media data. Further in-depth studies are needed to identify the determinants of social media posts described in this exploratory analysis.

**Keywords:** Seasonal allergic rhinitis; Social media; Twitter; Causal relationship; Infoveillance; Disease surveillance.

**Introduction**

The rapid growth of the internet has been accompanied by an increase in the use of social media data (from sources such as Twitter and Facebook) to explore and understand various phenomena. This form of social media monitoring can facilitate the effective analysis of large quantities of social media data produced in real time.

Large-scale quantitative analyses have been conducted using health-related social media data [1, 2], and the use of these data for disease surveillance (referred to as “infoveillance”) is gaining interest [3]. In particular, major advances have been made in the use of social media data to track the prevalence and spread of infectious diseases and other conditions. Among the infectious diseases targeted for surveillance, researchers have most actively applied these data to influenza surveillance [4–9]. Surveillance studies have also been conducted on enterohemorrhagic Escherichia coli outbreaks [10] and dengue fever [11]. In addition, social media data have been used to improve our understanding of Ebola [12] and Zika virus infections [13–15]. While disease surveillance efforts tend to focus on acute infectious diseases, studies have also been conducted on chronic diseases such as cancer [16], hypertension [16], asthma [16–18], diabetes [19], and seasonal allergic rhinitis (SAR) [20–22]. Systematic reviews have also been conducted on disease surveillance based on social media data [23–25].

These studies have contributed to public health by demonstrating moderately high correlations between fluctuations in the number of relevant social media posts and fluctuations in the number of patients for a specific disease. Public health authorities have also begun to adopt and apply currently available tools that use social media for influenza surveillance, such as HealthTweets.org [26, 27], Sickweather [28], and Now Trending [29]. Despite the reported correlations between actual disease prevalence and social media posts, the mechanism behind this relationship is poorly understood. In other words, the causal relationship between disease occurrence and the behavior of social media users remains unclear. For example, some individuals may only begin posting on social media after an existing condition becomes more severe. In contrast, others may have begun posting while having only mild symptoms, and seek medical treatment after their conditions worsen. Due to the
presence of these individual-level variations, the relationship between the numbers of social media posts and patient numbers remains inconclusive.

The lack of understanding of this relationship may have reduced the perceived reliability of disease surveillance based on social media data, as exemplified by the failure of Google Flu Trends [30]. This web service consistently overestimated influenza prevalence during the 2012–2013 season by over 50%, which led to a precipitous decline in its credibility as a surveillance system. The overestimations may have been influenced by the unusually early start of the 2012–2013 influenza season, which made it a frequent topic of discussion in many media outlets. As a consequence, this may have increased the number of people searching for influenza-related topics on Google. While the system algorithm was eventually updated, Google Flu Trends was shut down in 2015. In order to increase the credibility of disease surveillance based on social media data, there is a need to determine if there is a causal relationship between the behavior of social media users and the actual number of patients for a target disease.

SAR is an allergic disease that is so widely prevalent in Japan that it can be considered a national affliction. In particular, a large number of patients have been suffered from SAR induced by Japanese cedar pollen between February and April each year [33-35]. Although pollen is the main cause of SAR, symptoms only occur if a person is exposed to a quantity of pollen that exceeds that individual’s threshold level. As a result, there is no strong association between pollen count and patient numbers. In addition, SAR in Japan can also be triggered by pollen from other plant species (e.g., hinoki cypress and common ragweed), which complicates disease surveillance. However, the predicted prevalence of SAR in Japan is currently based solely on Japanese cedar pollen counts.

This study aimed to clarify the causal relationships among SAR prevalence, the behavior of social media users, and their actual actions in the real world. For example, we examined if more patients sought care after increases in pollen count, if Twitter users tweeted more after visiting a healthcare institute, and if patients visited hospitals more after SAR received increased attention on social media. The results of this analysis may support the use of social media in SAR surveillance, and shed light on the previously unknown behavior of SAR patients. In addition, we also discuss the content of some tweet examples.

Methods

Data
This analysis was conducted using the following datasets of pollen count, tweets, and SAR patient numbers as shown in Figure 1.

Figure 1. Time-based changes in data in Kanagawa Prefecture, Japan. The x-axes represent the date and the y-axes are the data counts for the following variables: (a) Changes in pollen count (mean daily pollen count from three
observation sites within Kanagawa Prefecture); (b) Changes in the number of SAR-related tweets; and (c) Changes in the number of SAR patients (daily number of patients from three participating healthcare institutes within Kanagawa Prefecture). The solid line represents the changes in the number of patients in which non-consultation days (i.e., days when an institute is closed) are supplemented by the patient numbers from the preceding and proceeding days; the dashed line represents the changes in the reported number of patients. Our analysis used the supplemented patient numbers.

**Pollen Count**

We examined daily pollen counts for Japanese cedar (the major cause of SAR in Japan) and hinoki cypress (which commonly complicates SAR). Japanese cedar and hinoki cypress pollen counts are measured hourly by automatic pollen counters located nationwide. Each prefecture has two to three of these devices, which are placed in urban areas with high population densities and in mountainous regions that are the major source of pollen production. Each prefecture’s mean daily pollen count is calculated based on aggregated hourly counts from multiple observation sites. For this study, we focused on Kanagawa Prefecture in the Kantō region of Japan. Hourly pollen counts from February 1 to May 31, 2017 were obtained from the Japanese Ministry of the Environment’s pollen observation system (designated “Hanakosan”) for analysis [31,32]. The data were recorded at three observation sites within Kanagawa Prefecture (the Second Annex of the Kanagawa Prefectural Government Building, the Kawasaki Life Science & Environment Research Center, and the Kanagawa Environmental Research Center), and the mean daily pollen counts across these three sites were calculated. Figure 1a shows the changes in mean daily pollen count in Kanagawa Prefecture from February to May 2017.

**SRA-related Tweets Numbers**

In an analysis of Twitter posts, we calculated the daily number of tweets that included the keyword “kafunshō” (or SAR). From among these tweets, we identified Twitter user profiles with location information (such as an area of residence or latitude/longitude data); the tweets were then classified according to the prefecture. There were 185,538 tweets from Japan that contained the target keyword between January 1 and May 31, 2017. Figure 1b shows the changes in the daily number of relevant tweets in Kanagawa Prefecture during this period.

**SRA Patient Numbers**

We analyzed the daily number of patients diagnosed with SAR during the period from January 1 to May 31, 2017. Although daily patient numbers are not generally publicized, we obtained these data from three healthcare institutes that agreed to participate in the study. These institutes were Sasaki Hospital (Yokohama City), Kawasaki Saiwai Clinic (Kawasaki City), and Kosugi ENT Clinic (Kawasaki City). Figure 1c shows the changes in the daily number of SAR patients that visited the participating healthcare institutes from January to May 2017. The dashed line represents the changes in the reported number of patients, and the solid line represents the changes in the number of patients in which non-consultation days
(i.e., days when an institute is closed) were supplemented by the patient numbers from the preceding and proceeding days.

**Analysis**

We sought to examine the causal relationships among the level of attention focused on SAR on Twitter (number of tweets), pollen count, and the number of SAR patients. While pollen counts may directly affect the number of tweets and SAR patients, the reverse is highly unlikely. As a result, we did not analyze the effects of tweet numbers and patient numbers on pollen count. Because this analysis used a data-driven approach, we employed the Granger causality test. This statistical hypothesis test determines if a particular time series is predictive of another time series.

To determine if time-variant factors affect the causal relationships among SAR prevalence, the behavior of social media users, and their actual actions in the real world; we conducted additional analyses where the study period (February to May 2017) was divided into two phases. The first was the main SAR phase, which generally occurs from February to April in the Kantō region for Japanese cedar-induced SAR [33]. The second was the concluding phase of the season, which generally occurs in May in the Kantō region for Japanese cedar-induced SAR. The data were analyzed as a differential time series of the differences between each day and the preceding day.

**Ethics statement**

This study did not require the participants to be involved in any physical and/or mental intervention. The participants’ information was unlinkable, anonymized, and de-identified prior to analysis. As this research did not use personally identifiable information, it was exempt from institutional review board approval in accordance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects stipulated by the Japanese national government.

**Results**

In addition to analyzing the overall study period (February to May 2017), the Granger causality test was also applied to the main SAR phase (February to April 2017). The results of the overall study period are presented in Table 1 and Figure 2a, and the results of the main SAR phase are presented in Table 1 and Figure 2b.

| Table 1. P values obtained by Granger causality test for pollen count, number of SAR-related tweets, and number of SAR patients. |
|-----------------|-----------------|-----------------|
| **Overall study period (February to May 2017)** | | |
| **Cause \ Effect** | **No. of tweets** | **No. of patients** |
| Pollen count | .04 | .04 |
| No. of tweets | - | .89 |
| No. of patients | .24 | - |
| **Main SAR phase (February to April 2017)** | | |
| **Cause \ Effect** | **No. of tweets** | **No. of patients** |
| Pollen count | .05 | .01 |
| No. of tweets | - | .02 |
| No. of patients | .47 | - |
Figure 2. Causal relationships between pollen count, number of SAR-related tweets, and number of SAR patients (Granger causality test results) for (a) the overall study period and (b) the main SAR phase. The Granger causality test did not reveal any causal relationships between these variables in the concluding phase of the season.

Figure 3. Peak number of SAR-related tweets, number of SAR patients, pollen count, and temperature. The x-axes represent the date and the y-axes are the min-max normalized data values for the four variables: The blue line represents the number of tweets, the green line represents the number of patients, the orange line represents the mean pollen count, and the grey line represents the mean temperature. (a) Peak number of tweets (Peak 1: January 30, 2017) and (b) Peak number of patients (Peak 2: March 1, 2017) and pollen count (Peak 3: March 7, 2017)

Effect of pollen count on SAR-related tweet numbers
As shown in Table 1 and Figure 2a, the Granger causality test rejected the null hypothesis that pollen count has no effect on SAR-related tweet numbers during the overall study period ($P = .04$); this indicates that pollen count has a causal effect on SAR-related tweet numbers. In contrast, the test did not reject this null hypothesis in the main SAR phase ($P = .05$), as shown in Table 1 and Figure 2b. We were unable to apply the Granger causality test to the concluding phase (May 2017) for these two variables.

Effect of pollen count on patient numbers
As shown in Table 1 and Figure 2a, the Granger causality test rejected the null hypothesis that pollen count has no effect on patient numbers during the overall study period ($P = .04$); this indicates that pollen count has a causal effect on SAR-related tweet numbers. In addition, the test also rejected the null hypothesis in the main SAR phase ($P = .01$), as shown in Table 1 and Figure 2b. We were unable to apply the Granger causality test to the concluding phase (May 2017) for these two variables.

Effect of SAR-related tweet numbers on patient numbers
As shown in Table 1 and Figure 2a, the Granger causality test did not reject the null hypothesis that tweet numbers have no effect on patient numbers during the overall study period ($P = .89$). In contrast, the test rejected this null hypothesis in the main SAR phase ($P = .02$), as shown in Table 1 and Figure 2b; this indicates that SAR-related tweet numbers have a causal effect on patient numbers during this phase. We were unable to apply the Granger causality test to the concluding phase (May 2017) for these two variables.

Effect of patient numbers on SAR-related tweet numbers
In both the overall study period and the main SAR phase, the Granger causality test did not reject the null hypothesis that patient numbers have no effect on SAR-related tweet numbers ($P = .24$ and $P = .47$, respectively). We were unable to apply the Granger causality test to the concluding phase (May 2017) for these two variables.
Discussion

In this analysis of the 2017 Japanese cedar pollen-induced SAR season, our results indicated that the level of attention focused on SAR on Twitter and pollen count may be able to predict the number of SAR patients. In addition, we examined the content of SAR-related tweets posted during the study period (January to May 2017) to gain further insight into these relationships.

The tweets in Textbox 1 are examples that were posted on January 30, 2017, which had the highest number of SAR-related tweets before the pollen count and SAR patient numbers increased. Pollen count is thought to increase in response to an increase in temperature and a decrease in humidity. In accordance with expectations, there was indeed a sudden increase in ambient temperature on January 30, 2017 (Fig. 3a), which may have caused more sensitive users to identify and report symptoms ascribed to SAR. This in turn may have led to the high level of attention on Twitter on this day.


I’ve been down with hay fever since that windy day. I don’t need this lol. Pollen is here, isn’t it? My hay fever’s not that bad, but I could feel it’s “arrival” 2/3 days ago. This morning I was hard-core sneezing and my nose can’t stop running...
Looks like it’s hay fever (‘Д ‘)
Good morning (^-^)/ I keep hearing about hay fever these few days...everyone’s most hated season is coming again, eh? You gotta eat lotus roots! Have a great day, everyone!
Uh oh. I haven’t got any tissues. I’m dying. My runny nose won’t stop. I’m about to have hay fever. This is gonna be rough.
I’m totally convinced that once it starts to get warm, then hay fever comes along.
I haven’t got hay fever, but I can kind of feel the pollen flying.
When it suddenly gets warmer I can feel the hay fever coming
I don’t know if its hay fever or just the temperature differences in winter...but I’m snuffling.
Today, it’s warm and windy, and I’m sneezing lots — is the pollen flying? I also heard that this year’s pollen count is 4.4 times more than last year’s. I already had a pretty rough time last year with my severe hay fever...stuffy nose, itchy eyes...I hate this. During this period I’d like to rip out my nose and eyeballs. Seriously.
I hate this my nose won’t stop running. I don’t even know if this is the remnants of a cold or the effects of hay fever for real lol.
I kind of feel like this year’s hay fever is already starting. I’ll stock up on OTC meds tomorrow (T ^ T)
It...it’s finally here (>_<)... hay fever (; >_ < ;) my eyes are itching and my nose is running and my head is heavy...(>_<) it’s usually empty tho (^ ^
My face is suddenly swollen today, and my eyes are bleary (⊙ _ ⊙;) it’s hay fever!!!
I’m home! maybe it’s the warm weather, but I feel a bit hay fever-ish... what a pain
When it’s this warm, hay fever’s here---- ; _ ; #nhk11
Next, the tweets in Textbox 2 are examples that were posted on March 1, 2017, which had the highest number of SAR patients seeking care at the three participating healthcare institutes during the study period. Coincidentally, there was an extremely low pollen count on this day. While there appeared to be a slight decrease in tweet numbers (Fig. 3b), these tweets included those from SAR patients who were going or had gone to seek treatment.

Textbox 2. Examples of tweets posted on March 1, 2017.

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Created on 1st March, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>I'm fine when I'm outside, but I suddenly get hay fever symptoms around midnight. What's up with this delayed attack.</td>
<td></td>
</tr>
<tr>
<td>I woke up cos I couldn't breathe cos my nose was stuffed cos of hay fever (&gt;-&lt;)</td>
<td></td>
</tr>
<tr>
<td>My sleeping time and concentration are dwindling away because of hay fever</td>
<td></td>
</tr>
<tr>
<td>Argh. This hay fever headache is massive...</td>
<td></td>
</tr>
<tr>
<td>OH NO. I forgot to take my hay fever meds (-_-;)</td>
<td></td>
</tr>
<tr>
<td>My eyes are totally red cos of hay fever (☝️ ՞️️ ۝️ ՞️️ ☝️) But I still wanna use my colored contacts ~ (☝️ ՞️️ ۝️️ ՞️️)☝️</td>
<td></td>
</tr>
<tr>
<td>I just took hay fever meds. I'll sleep a bit more. I dunno why my shoulders are so stiff lately.</td>
<td></td>
</tr>
<tr>
<td>This anti-hay fever mask has NO effect LOL</td>
<td></td>
</tr>
<tr>
<td>I forgot my eye drops and nasal spray for hay fever but I haven't got any symptoms.</td>
<td></td>
</tr>
<tr>
<td>Maybe the non-drowsy oral meds are enough? Or is the real deal still to come?</td>
<td></td>
</tr>
<tr>
<td>I'm using a mask but my nose is running from hay fever........</td>
<td></td>
</tr>
<tr>
<td>Came to see the doctor for hay fever before work but it's really crowded</td>
<td></td>
</tr>
<tr>
<td>I'm here at the ENT. The doctor recommended an actual treatment to cure hay fever and not just suppress the symptoms but it looks like a pain to keep coming here. It seems I have to come here every month for 5 years, even in the off season. I want to do it if I have time next year!!</td>
<td></td>
</tr>
<tr>
<td>So this is hay fever. If I go outside without a mask my nose becomes a waterfall...</td>
<td></td>
</tr>
<tr>
<td>When I woke up today I was sniffing more than usual and my throat hurts... (&gt;-&lt;)...is this hay fever or a cold...if this continues tomorrow, I'll probably come back tomorrow in even worse condition (^◇^;) last time I thought it was a cold and felt pretty confident, but it got worse and I lost my voice LOL</td>
<td></td>
</tr>
</tbody>
</table>

Finally, the tweets in Textbox 3 are examples that were posted on March 7, 2017, which had the highest pollen count during the 2017 SAR season. As shown in Figure 3b, there was a sudden increase in pollen count for several days before peaking on this day. While there was a reduction in the number of tweets, these tweets included people who had SAR symptoms for the first time this season, as well as reports of the worst symptoms for this season. There was also an increase in the number of SAR patients, and we confirmed that there were tweets where patients reported seeking treatment at healthcare institutes.


<table>
<thead>
<tr>
<th>Tweet</th>
<th>Created on 7th March, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hay fever sucks</td>
<td></td>
</tr>
<tr>
<td>Good morning☆Hay fever has arrived^_^- my eyes and nose are so itchy~ I'll try to be</td>
<td></td>
</tr>
</tbody>
</table>
cheerful today as well(^o^)/
I've run out of hay fever meds, so if I seem to keep sneezing, I'll probably have to get
more meds from the doc.
Oh man, my face is swollen and painful. Hay fever? Or maybe allergies?
I've got hinoki hay fever and dust sensitivity, so I've GOT to have a mask for this old
house in Yamanashi.
My hay fever is going completely crazy today. Forgetting the tissues was a fatal
mistake.
Sudden spike in the number of people with hay fever symptoms!
I thought I was sick, but the hospital told me I had hay fever — unbelievable...
My hay fever is horrible today. This is the worst for this year (T_T)
I just came back from the hospital, and I've gotta go again...my hay fever debut orz
I was totally phased out with hay fever and almost forgot to go to school.
My nose won't stop running. I think it's hay fever.
My hay fever is crazy so I've got shades and a mask on. Totes feel like a celeb.
Today's hay fever is really rough. I'm completely done in with this triple play of
sneezing, runny nose, and itchy eyes.
The hay fever eye drops work only for just over 10 minutes. I really got to get some oral
meds.
I haven't even stepped out of the house but the hay fever is so bad I don't feel like doing
anything (°_°)
I'm sleepy already, this could also be the hay fever at work
Hay fever where you don't stop sneezing is a real pain ( 'ω' )
It's not a cold, but my head hurts, and when that's settling I've got a runny nose, am
sneezing and coughing all at once. This is the start of legit hay fever ( _´)ゞ
Been a while since I've had this full-blown hay fever. I blew my nose and it started
bleeding, so I spent almost the whole day with a nosebleed. It won't stop.
The inside of my mouth itches cos of hay fever

Through our analysis of pollen count and tweet numbers, we were able to observe
the difference in data characteristics between these variables. As shown in Figure
1a, pollen count is affected by external factors (such as temperature, rainfall, wind
speed, and wind direction), and there are substantial fluctuations throughout the
study period. Accordingly, it is difficult to predict at a glance when the pollen season
would end. On the other hand, the number of tweets showed some fluctuations (Fig.
1b), but there was a general increase from January that peaked in February and
March, and a steady downward trend was observed thereafter. In this way, we were
able to visualize the trend toward the conclusion of the SAR season. Our analysis
showed that the combined use of data with different characteristics not only
provided information on SAR prevalence but also enabled observation of the SAR
season as it progressed.

The development of an analytical model that accounts for the different
characteristics of the datasets while providing insight into the causal relationships
may enable highly reliable disease surveillance.
Limitations
Further in-depth studies are needed to identify the determinants of social media posts described in this exploratory analysis. In addition, this study focused on one SAR season, and downstream studies should be conducted using data that are continuously collected over multiple seasons. Finally, there is a need to conduct intervention-based prospective studies to gain a more accurate understanding of the causal relationships in these variables.

Conclusions
While social media data are increasingly used in disease surveillance, there is a need to improve the credibility of these surveillance systems in order to promote their implementation and acceptance in society. Understanding the causal relationships between the behavior of social media users and actual patient numbers is an important step to increase the credibility of these surveillance systems. In this study, we analyzed data on pollen count, the number of tweets, and the number of patients during the 2017 SAR season in Japan using the Granger causality test, and shed light on the causal relationships among these variables. Increases in pollen count were found to cause increases in the number of tweets and patients. In addition, increases in social media posts (i.e., tweets) also caused increases in patient numbers. The main SAR phase and the concluding phase of the season appeared to have different characteristics. Accordingly, disease surveillance based on social media data should be adjusted to account for these time-based differences.

Acknowledgements
This work was supported in part by an ACT-I research grant from the Japan Science and Technology Agency and the Research Program on Emerging and Re-emerging Infectious Diseases from Japan Agency for Medical Research and development, AMED (#16768699). The authors appreciate the advice and cooperation of the Japan Anti-Pollinosis Council and NPO Therapeutic Promotion Society for Pollinosis and Rhinosinusitis for this study. The authors are also grateful to the physicians and staff of the three participating hospitals for providing the patient statistics.

Conflicts of Interest
None declared.

Abbreviations
SAR: Seasonal Allergic Rhinitis.

References


