Clinical Age-Specific Seasonal Conjunctivitis Patterns and Their Online Detection in Twitter, Blog, Forum, and Comment Social Media Posts

Michael S. Deiner,[1,2] Stephen D. McLeod,[1,2] James Chodosh,[3] Catherine E. Oldenburg,[1,2,4] Cherie A. Fathy,[5] Thomas M. Lietman,[1,2,4] and Travis C. Porco[1,2,4]

[1]Francis I. Proctor Foundation for Research in Ophthalmology, University of California, San Francisco, San Francisco, California, United States
[2]Department of Ophthalmology, University of California, San Francisco, San Francisco, California, United States
[3]Massachusetts Eye and Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, United States
[4]Department of Epidemiology and Biostatistics, Global Health Sciences, University of California San Francisco, San Francisco, California, United States
[5]Beth Israel Deaconess Medical Center/Brockton Signature Hospital, Brockton, Massachusetts, United States

Correspondence: Travis C. Porco, Francis I. Proctor Foundation, University of California, San Francisco, 513 Parnassus, San Francisco, CA 94143, USA; travis.porco@ucsf.edu.

Submitted: August 16, 2017
Accepted: January 5, 2018

Citation: Deiner MS, McLeod SD, Chodosh J, et al. Clinical age-specific seasonal conjunctivitis patterns and their online detection in twitter, blog, forum, and comment social media posts. *Invest Ophthalmol Vis Sci.* 2018;59:910–920. https://doi.org/10.1167/iovs.17-22818

Purpose. We sought to determine whether big data from social media might reveal seasonal trends of conjunctivitis, most forms of which are nonreportable.

Methods. Social media posts (from Twitter, and from online forums and blogs) were classified by age and by conjunctivitis type (allergic or infectious) using Boolean and machine learning methods. Based on spline smoothing, we estimated the circular mean occurrence time (a measure of central tendency for occurrence) and the circular variance (a measure of uniformity of occurrence throughout the year, providing an index of seasonality). Clinical records from a large tertiary care provider were analyzed in a similar way for comparison.

Results. Social media posts machine-coded as being related to infectious conjunctivitis showed similar times of occurrence and degree of seasonality to clinical infectious cases, and likewise for machine-coded allergic conjunctivitis posts compared to clinical allergic cases. Allergic conjunctivitis showed a distinctively different seasonal pattern than infectious conjunctivitis, with a mean occurrence time later in the spring. Infectious conjunctivitis for children showed markedly greater seasonality than for adults, though the occurrence times were similar; no such difference for allergic conjunctivitis was seen.

Conclusions. Social media posts broadly track the seasonal occurrence of allergic and infectious conjunctivitis, and may be a useful supplement for epidemiologic monitoring.

Keywords: infectious conjunctivitis, allergic conjunctivitis, machine learning, social media, Twitter

Using "big data" to complement traditional clinical case reporting has been conducted for a number of conditions,[1–16] including conjunctivitis.[17–21] It has been argued that epidemiologic application of such data may result in bias,[22,23] but could also provide information not otherwise available.[24] Aside from neonatal cases, no CDC reporting system for conjunctivitis exists, despite having some economic[25] and public health[26] importance. We asked whether social media streams reveal important features of the epidemiology of conjunctivitis, such as age-specific seasonal occurrence, for instance.

Previously, we found that temporal patterns of conjunctivitis, whether allergic or infectious, correlated with searches and tweets.[20] In this paper, using data from a longer time period, we test whether the seasonality of clinical case counts for both infectious and allergic conjunctivitis is related to the seasonality of social media posts. In addition to tweets, we examine posts and comments on blogs and Internet forums.

Methods

EMR Clinical Data Acquisition

All UCSF electronic medical records (EMR) from June 3, 2012, to July 16, 2016, were queried. Queries were constructed to identify cases of: (1) conjunctivitis (all encounters with a diagnosis name containing "conjunctivi") and controls; (2) glaucoma (all encounters with a diagnosis name containing "glaucoma"); (3) macular degeneration (ICD9 diagnosis code of 362.5 and ICD10 equivalents); (4) corneal ulcer (ICD9 diagnosis code of 370.x and ICD10 equivalents); (5) common cold (ICD9 diagnosis code of 460.x and ICD10 equivalents); and (6) influenza (ICD9 level 3 group of "Influenza" and ICD10 equivalents). Using ICD9/10 codes and diagnosis names, conjunctivitis was further classified as infectious conjunctivitis, allergic conjunctivitis, or other conjunctivitis (see Supplementary Material for details). These records included diagnoses from all provider specialties. For each diagnosis, a maximum of

Table 1. Distributions of EMR and Social Media Post Disease Groups, by Conjunctivitis Disease Group and Age

Data Source	Age Category, y	Infectious Conjunctivitis, *n* (%)	Allergic Conjunctivitis, *n* (%)	Ratio	Other Conjunctivitis, *n* (%)
Clinical	0–5	1,685 (27)	127 (5)	13.3	72 (14)
	6–17	1,099 (18)	581 (22)	2.1	42 (8)
	18–39	1,165 (19)	566 (22)	1.9	158 (31)
	40+	2,237 (36)	1,340 (51)	1.7	241 (47)
	All ages	*6,186*	*2,614*	*2.4*	*513*
Twitter	Younger	8,745 (5)	1,553 (1)	5.6	–
	Older	174,556 (95)	230,499 (99)	0.8	–
	All ages	*183,301*	*232,052*	*0.8*	–
Forums	Younger	4,785 (31)	1,780 (10)	2.7	–
	Older	10,649 (69)	16,337 (90)	0.7	–
	All Ages	*15,434*	*18,117*	*0.9*	–

Percentages are shown within each disease group (*column*). The ratio column shows the proportion of infectious to allergic conjunctivitis, by age.

only one encounter per day per patient was included for subsequent analysis. Weekly counts were grouped by overall diagnosis category. Infectious conjunctivitis, allergic conjunctivitis, other conjunctivitis and flu were also tabulated in two ways: most common top provider specialties and age categories (0–5 years, 6–17 years, 18–39 years, and 40 or more years of age).

Social Media Data Acquisition

Using a commercial social media analytics platform Crimson Hexagon,[27,28] we identified social media posts related to a case of a person (or group of people) having any form of infectious or allergic conjunctivitis or eye allergy using a Boolean query (to refine and limit results using keywords and phrases with the operators AND, OR, NOT). Cinematic or cultural references to pink eye were similarly excluded. To further refine the data into distinct infectious or allergic groups, we then trained the crimson Hexagon BrightView classifier to classify posts as either "infectious conjunctivitis" (posts concerning conjunctivitis or pink eye but not related to allergy); "allergic conjunctivitis" (posts concerning eye allergy or eyes with allergic conjunctivitis symptoms such as itching, but not related to infectious spread); and "irrelevant/uncertain" (posts concerning cosmetics, humor, other disease, emotional crying, or otherwise unclassifiable). The BrightView classifier is a supervised learning algorithm based in part on stacked regression analysis of simplified numerical representation of text.[29] Posts that may potentially address infectious or allergic conjunctivitis were collected as a corpus of tweets from Twitter and a separate corpus of pooled posts from forums, blogs, and public comment sections ("forums"). All forums, blogs, and comment sections available through Crimson Hexagon were queried. Within the identified infectious and allergic conjunctivitis Twitter and forums posts, a Boolean classification was then used to characterize posts as related to young people or children, or not being related to young people or children, in order to create "younger" and "older" age-based subgroups. Boolean queries above were also used to remove posts containing common terms, phrases, song names, user accounts, and so forth, if obviously not related to posts about occurrence of known conjunctivitis in a person (see Appendix and Supplementary Material for more query details).

Validation of Machine-Coded Classification

Validation of the machine-coded classifications of infectious and allergic conjunctivitis was assessed by comparison with human classifications. A random sample of the classified posts, excluding any used in training the machine, was selected. Two independent human raters were masked to the machine-coded classifications and to each other's classifications. Humans rated posts independently, and then a modified Delphi process was used to arrive at a consensus human rating. The consensus human rating was then compared to the BrightView classifier using percent agreement. Details are provided in the Appendix.

Statistical Approach

Clinical and social media data weekly counts were analyzed using negative binomial regression[30] with a cubic polynomial in time to account for secular trends, and cyclic cubic splines[31] as a flexible model of arbitrary seasonal variation (with eight knots per year). Estimation was conducted by maximum likelihood, yielding detrended and smoothed estimates of the average count over the course of 1 year. Based on these estimates, we then computed the circular mean, the circular variance, and the relative amplitude. The circular mean provides a measure of central tendency for estimated occurrence frequency over a year (e.g., the circular mean week for school graduations in the United States might be the first week of June). The circular variance measures the degree of uniformity of seasonal occurrence, ranging from 0 when all events occur at the same instant every year, and 1 when events are uniformly distributed throughout the year and no seasonal variation exists. Hypothesis testing was conducted using the likelihood ratio χ^2. We also compared the similarity of time series using Spearman rank correlation by first detrending (using negative binomial regression with a cubic polynomial in time, and using time series bootstrap with a fixed window of length 8). When we found evidence of outliers in the time series, we conducted sensitivity analysis using additional predictors of the form $-(t - t_0)(t - t_1)$ (where t_0 and t_1 are the beginning and end of each sequence of outliers). All computations were conducted in R, version 3.3.1 for Macintosh (R Foundation for Statistical Computing, Vienna, Austria). Details are provided in the Appendix.

IRB Approval

This study followed the tenets of the Declaration of Helsinki, with approval obtained from the UCSF institutional review board prior to commencing this study (IRB approval# 14-14743).

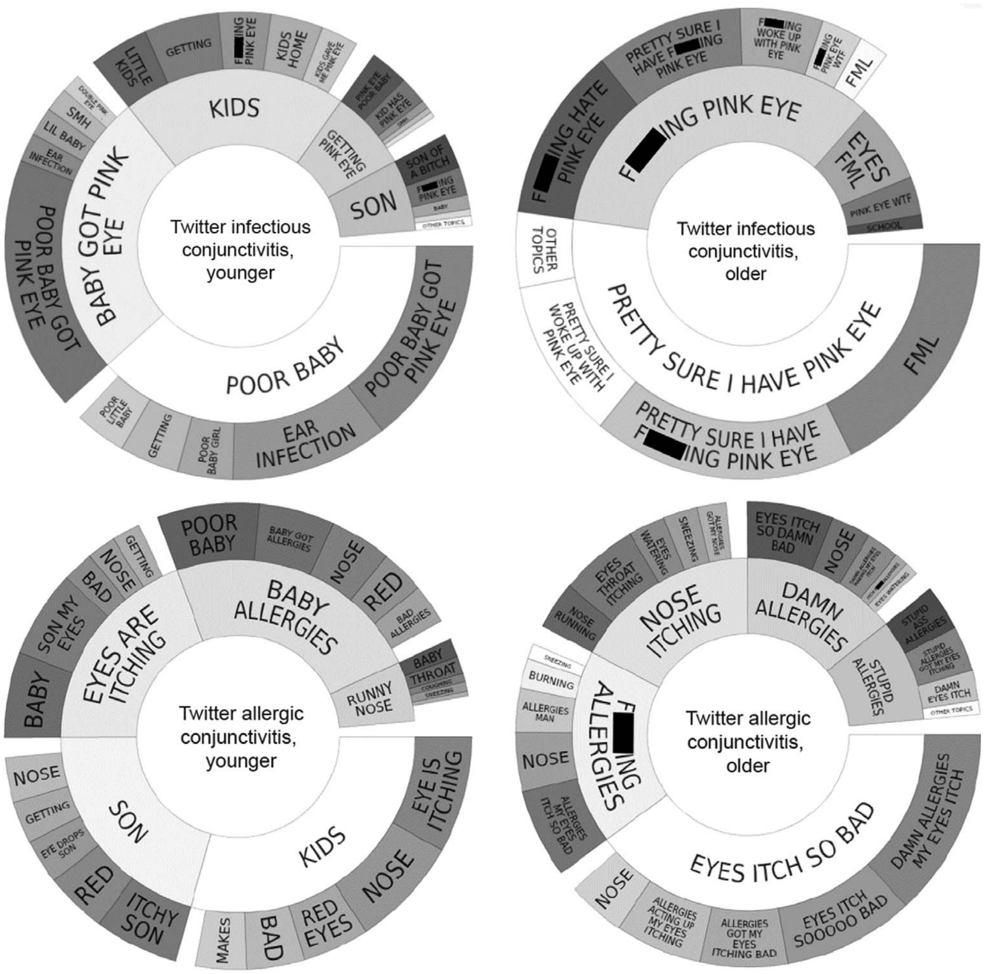

Figure 1. Conversation topics, Twitter. Topic wheels created from a random sample of 1000 posts used for analysis. *Left*: younger. *Right*: older. *Top*: infectious. *Bottom*: allergic. Source: Crimson Hexagon.

Results

We examined the period June 3, 2012, to July 16, 2016. Distributions of clinical conjunctivitis diagnosis groups for all ages combined, as well as for age-based subgroups, are shown in Table 1. For the clinical data, we extracted encounters for infectious conjunctivitis, allergic conjunctivitis, and other conjunctivitis cases (such as those resulting from chemical exposure). For comparison, we also included five additional conditions, including influenza. The EMR patient data set yielded 108,699 total patient records for these eight conditions. Counting each patient only once for each category, a total of 31,037 patients were analyzed. Of these, we analyzed 4713 allergic conjunctivitis visits for 2614 patients, 8036 infectious conjunctivitis visits for 6186 patients, and 810 other conjunctivitis visits for 513 patients. For influenza, we analyzed 3237 visits for 2270 patients. Clinical infectious, allergic, and other conjunctivitis encounters had differing distributions depending on specialty. Of the total infectious conjunctivitis patient encounters, the most (27%) were seen in pediatrics followed by internal medicine at 22% and ophthalmology at 14%. Of the total allergic conjunctivitis patient encounters, pediatrics and ophthalmology ranks were reversed from infectious conjunctivitis: the most (32%) allergic encounters were seen in ophthalmology, followed by optometry at 17% and pediatrics at 14%. For other conjunctivitis patient encounters, the most (44%) were seen in ophthalmology, followed by optometry at 22% and pediatrics at 7%. Emergency medicine contributed 11% of infectious, 2% of allergic, and 6% of other conjunctivitis cases, with conjunctivitis-related sex, age, and seasonal results similar to that found for a national emergency medicine clinical database.[32]

The Twitter query resulted in 183,301 infectious conjunctivitis posts and 232,052 allergic conjunctivitis posts. The queried forums, blogs and comments ("forums") pooled results (consisting of approximately 89% forums, 9% blogs, and 2% comments), yielded 15,434 infectious conjunctivitis posts and 18,117 allergic conjunctivitis posts overall.

Validation of Social Media Data Machine-Coded Classifications

For Twitter, comparison of the consensus human classification to the machine-coded classifications showed that humans agreed 92% (95% confidence interval [CI]: 86%–96%) of the time with the machine-coded classifications, with human agreement 89% (95% CI: 79%–95%) of the time when the machine classification was infectious, and 95% (95% CI: 87%–99%) of the time when the machine classification was allergic. For Twitter posts where the consensus human classification was infectious conjunctivitis, the machine-coded classification was infectious 100% of the time (57/57) and allergic 0% of the time. For posts where the consensus human classification was

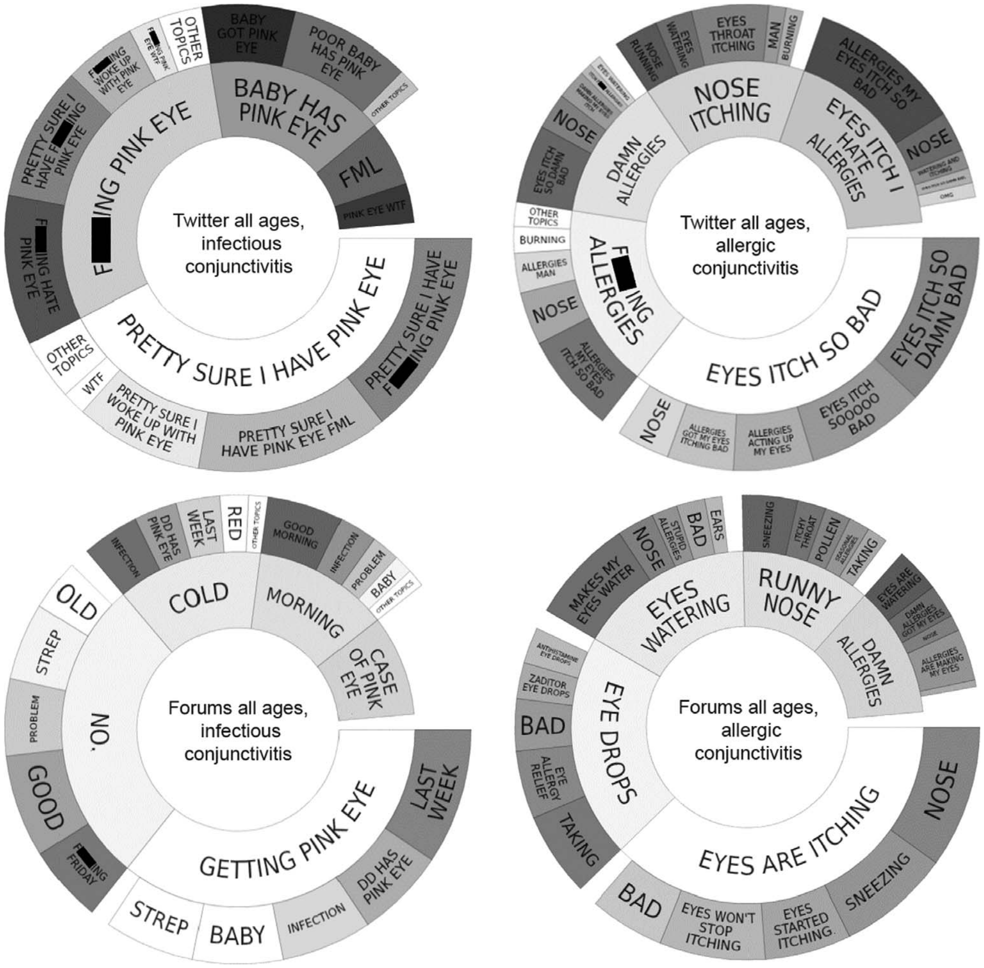

Figure 2. Conversation topics, all ages. *Left*: infectious conjunctivitis. *Right*: allergic. *Top*: Twitter. *Bottom*: forums. Source: Crimson Hexagon.

allergic conjunctivitis, the machine-coded classification was allergic 100% (61/61) of the time and infectious 0% of the time.

For forums posts, comparison of the consensus human classification to the machine-coded classifications showed that humans agreed 55% (95% CI: 42%–67%) of the time with the machine classifications, with human agreement 59% (95% CI: 41%–76%) of the time when the machine classification was infectious, and 50% (95% CI: 32%–68%) of the time when the machine classification was allergic. Only 1 case of the disagreement between human consensus and machine-coded classifications was due to opposing forms of human and machine-assigned conjunctivitis classifications. Almost all disagreement was because the consensus human classification found 43% of the machine-coded conjunctivitis forums posts to be uncertain or unrelated to conjunctivitis. For comparison, for Twitter posts, the consensus human classification found only 8% of the machine-coded conjunctivitis Twitter posts to be uncertain or unrelated to conjunctivitis. This may reflect the longer length and less uniform nature of forums posts. For forums posts where the consensus human classification was infectious conjunctivitis, the machine-coded classification was infectious 100% of the time (19/19) and allergic 0% of the time. For forums posts where the consensus human classification was allergic conjunctivitis, the machine-coded classification

Table 2. Seasonal Characteristics of Conjunctivitis in Clinical EMR and Social Media Posts

Data Source	Disease Group	Mean Week	Circular Variance	Relative Amplitude	*P* Value
Clinical	Infectious conjunctivitis	March 18 (March 10–26)	0.79 (0.76–0.82)	0.58 (0.53–0.65)	$P < 0.001$
Twitter	Infectious conjunctivitis	February 24 (February 17–March 3)	0.90 (0.89–0.91)	0.38 (0.35–0.42)	$P < 0.001$
Forums	Infectious conjunctivitis	March 3 (February 20–March 13)	0.80 (0.77–0.84)	0.63 (0.55–0.70)	$P < 0.001$
Clinical	Allergic conjunctivitis	May 30 (May 18–June 13)	0.85 (0.82–0.88)	0.57 (0.49–0.64)	$P < 0.001$
Twitter	Allergic conjunctivitis	May 8 (May 6–11)	0.74 (0.73–0.75)	0.73 (0.72–0.75)	$P < 0.001$
Forums	Allergic conjunctivitis	May 17 (May 8–27)	0.79 (0.75–0.82)	0.60 (0.56–0.69)	$P < 0.001$
Clinical	Other conjunctivitis	June 18 (May 24–July 14)	0.87 (0.82–0.92)	0.50 (0.42–0.67)	$P < 0.001$
Clinical	Influenza	February 1 (January 29–February 3)	0.33 (0.31–0.36)	0.97 (0.96–0.98)	$P < 0.001$
Clinical	Corneal ulcer	June 1 (April 20–July 6)	0.95 (0.93–0.98)	0.32 (0.26–0.44)	$P < 0.001$

The circular mean, circular variance, and (relative) amplitude are derived from estimated cyclic splines; the *P* value reflects the test of the null hypothesis of nonseasonality.

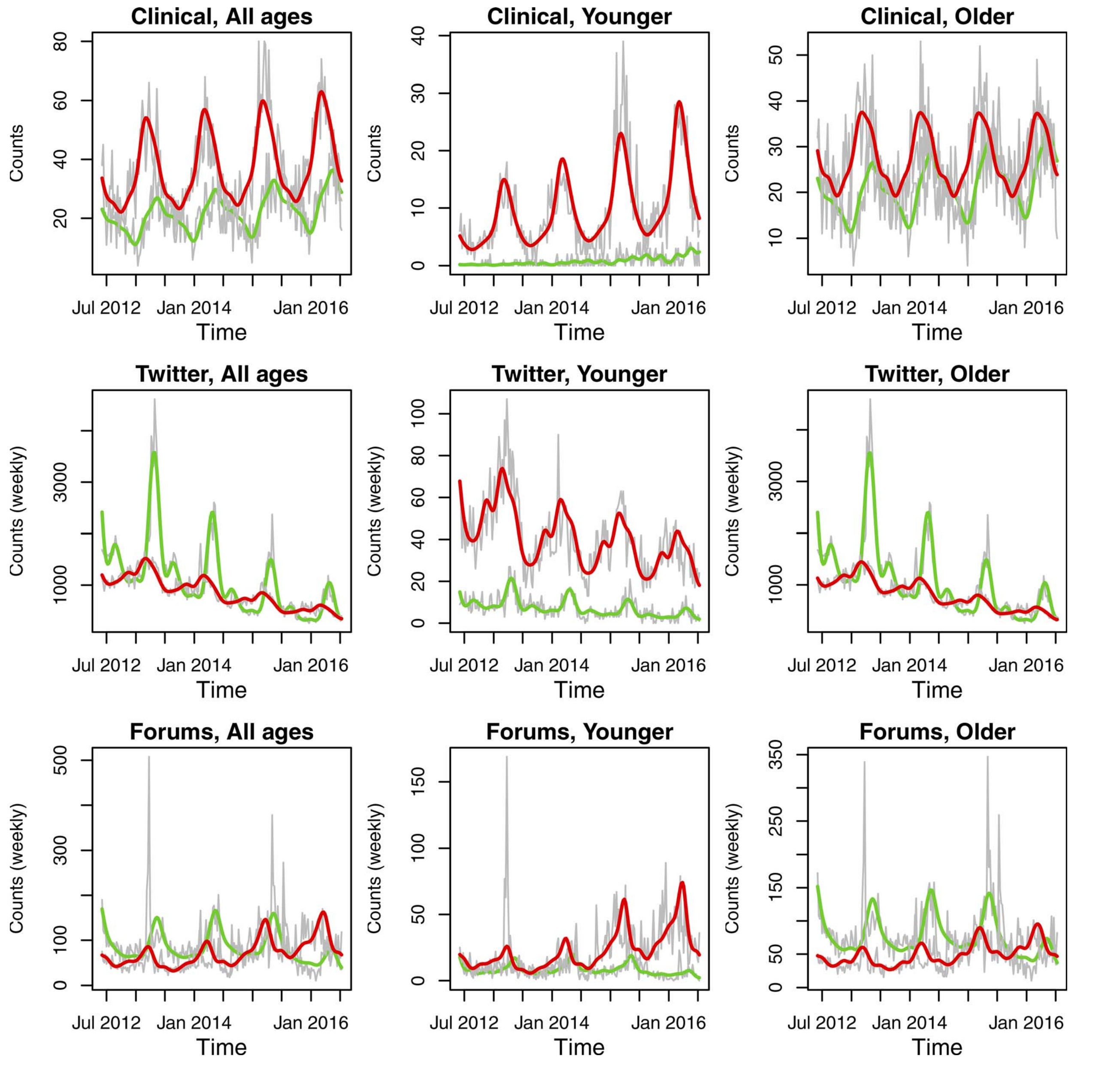

Figure 3. Weekly EMR and social media counts and estimated seasonal pattern fitted curves: conjunctivitis groups and age groups for clinical and social media data. Weekly count data, using data from over a multiple years, were analyzed using negative binomial regression to create fitted seasonal curves. *Panels* show raw weekly data and corresponding fitted curves, these can be compared between conjunctivitis infectious, allergic and other groups for both EMR clinical data as well as analogous Twitter or Forum post conjunctivitis post groups. *Rows*: Clinical cases (row 1); Twitter (row 2); and forums (row 3). *Columns*: All ages combined (column 1); younger ages (column 2); older ages (column 3). *Colors*: fitted curves for infectious conjunctivitis (*red*); fitted curves for allergic conjunctivitis (*green*); observed weekly counts (*gray*). Each *tick* on the *horizontal axis* represents 6 months and each date listed corresponds to the *tick mark* centered above its listed date.

was allergic 94% of the time (16/17) and infectious 6% of the time (1/17).

Age-Based Query Results

Table 1 summarizes age-based query results for EMR clinical, Twitter, and forum data. The ratio of infectious to allergic conjunctivitis posts by age subgroup demonstrates a strong similarity between clinical data, Twitter posts, and forum posts in that younger ages manifest a much higher ratio of infectious to allergic cases than older ages. Text content for conjunctivitis type and age-based subgroups of Twitter posts can be compared in the topic wheel visualizations shown in Figure 1, depicting apparent differences in some main topics and subtopic content for each subgroup.[33] Additional content comparison of Twitter and forum age-based subgroups for infectious conjunctivitis are shown in Figure 2.

Estimated Seasonal Patterns and Comparisons

Weekly infectious conjunctivitis cases, tweets, and posts are shown from 2012 to 2016 in Figure 3 (gray, smoothed curve in red). The figure also shows allergic conjunctivitis (gray, with smoothed curve in green). Detrending yielded estimated seasonal variation over the course of 1 year, shown in the smoothed curves in Figure 4. From these detrended smoothed

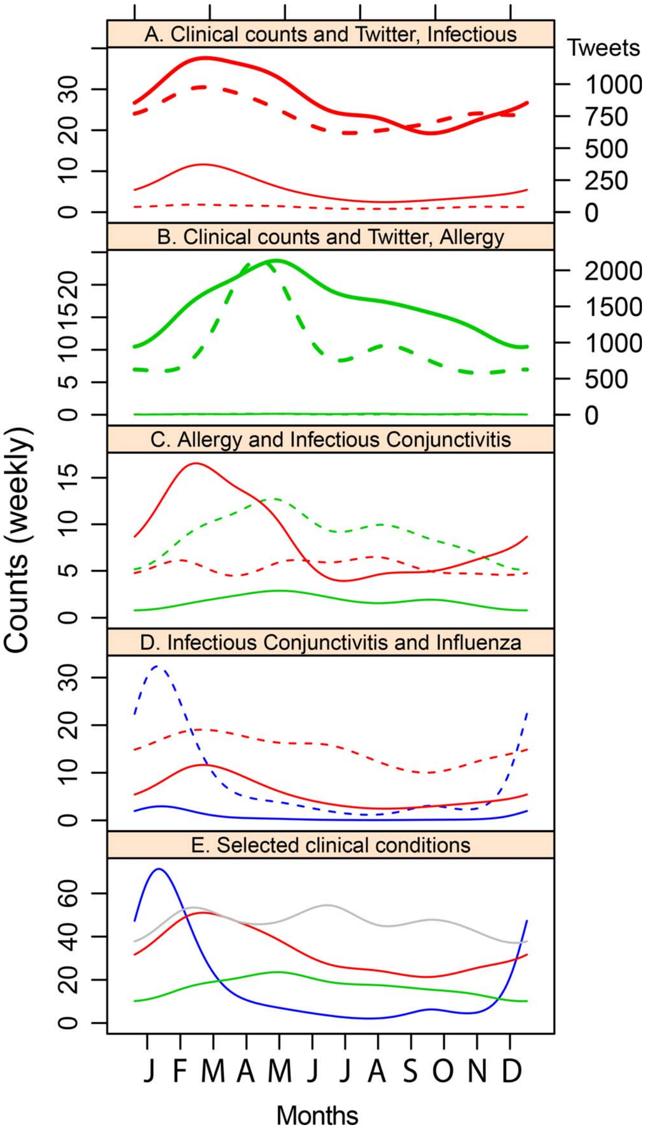

FIGURE 4. Smoothed detrended seasonal curves. (**A**) Infectious conjunctivitis for older (*thick lines*) and younger individuals (*thin lines*); clinical EMR (*solid lines*); and Twitter (*dashed lines*). (**B**) Allergic conjunctivitis for older (*thick lines*) and younger (*thin lines*); clinical EMR (*solid lines*); and Twitter (*dashed lines*). (**C**) Infectious conjunctivitis (*red*) and allergic conjunctivitis (*green*), for pediatrics (*solid lines*) and ophthalmology (*dashed lines*). (**D**) Clinical influenza (*blue*) and infectious conjunctivitis (*red*), for older (*solid lines*) and younger individuals (*dashed lines*). (**E**) Clinical influenza (*blue*); allergic conjunctivitis (*green*); infectious conjunctivitis (*red*); and corneal ulcers (*gray*), all ages. *X-axis*: *top tick marks* are 10-week intervals starting at week 0, *bottom tick marks* indicate middle of months.

curves, we calculated summary statistics representing specific seasonal features (i.e., the circular mean, circular variance, and the relative amplitude; shown in Tables 2, 3; Fig. 5). The estimated circular mean for infectious conjunctivitis clinical data was March 18 (95% CI: March 10–26), providing a measure of occurrence time within a year. The estimated circular variance was 0.79 (95% CI: 0.76–0.82), providing a measure of seasonality.

Overall, similar seasonal features were detected between social media and clinical data, for all ages combined. (Because seasonal occurrence data are angular data, we used the circular mean week to summarize the central tendency.) The circular mean week of infectious conjunctivitis was similar for data from EMR, Twitter, and forums; and the circular mean week of allergic conjunctivitis was similar for data from EMR, Twitter and forums (see Table 2). The circular mean week for infectious conjunctivitis preceded the allergic conjunctivitis mean week by approximately 10 weeks for all three data sources (see Table 2, Figs. 3, 4). Weekly infectious conjunctivitis counts for all three data sources were more strongly correlated with each other than with any allergic conjunctivitis data source (Spearman correlation, please see Table 4). Similarly, counts for allergic conjunctivitis for all three data sources were more strongly correlated with each other than with any infectious conjunctivitis data source (see Table 4).

For both infectious and allergic conjunctivitis, the seasonality of tweets and forum posts is similar to that of clinical conjunctivitis, in both younger and older age groups. Since the age classification for clinical cases is not the same as for tweets and forum posts, we do not expect identical seasonal patterns. However, the mean occurrence day of infectious conjunctivitis for the youngest age group for the clinical data was March 9 (95% CI: March 5–15). The mean occurrence day of tweets for infectious conjunctivitis for younger people was February 26 (95% CI: February 19–March 6); see Table 5 for details. Figure 5 shows the seasonality (one minus the circular variance) as a function of the circular mean occurrence time, by age and condition. Allergic conjunctivitis appears somewhat later in the year than infectious conjunctivitis. Among infectious conjunctivitis cases, children show greater seasonal variation than older individuals. These patterns are apparent in both the clinical data and in social media. Standard errors are given in Table 5. Broadly speaking, for infectious conjunctivitis in both the younger and the older age groups, tweets and forums are more Spearman correlated with infectious clinical cases than for allergic clinical cases (not shown).

Comparing seasonal infectious conjunctivitis features between age groups, for Twitter, forums, or EMR data, infectious conjunctivitis at younger ages had a higher degree of seasonality than at older ages (see Fig. 3, column 2 versus 3; Fig. 5 solid red versus open red). Besides using patient age groups, as an alternative approach to compare seasonal characteristics of younger clinical cohorts to older cohorts, we also compared seasonal characteristics of patients seen in pediatrics to other specialties and found that infectious conjunctivitis for pediatrics had a much higher degree of seasonality and earlier mean week than for ophthalmology (see Fig. 4C) as well as, to a lesser extent, for other clinical specialties (data not shown).

Comparing between infectious and allergic conjunctivitis by age groups, there were more infectious than allergic cases at younger ages than at older ages, for EMR and social media data (see Table 1 "Ratio", Fig. 3, columns 2–3). At younger ages, infectious conjunctivitis had stronger circular variance seasonality than allergic conjunctivitis, for EMR and forums (but not Twitter). Additionally, for all three data sources, at younger ages, infectious conjunctivitis had a larger relative amplitude than allergic conjunctivitis (see Table 5; Fig. 5, red closed versus green closed; Figs. 4A, 4B, thin lines). Inversely, at older ages, allergic conjunctivitis had equal or stronger circular variance seasonality and relative amplitude than infectious conjunctivitis did, for all three data sources (see Table 5; Fig. 5; Fig. 4A, 4B, thick lines).

Not all seasonal features, however, were the same between EMR, Twitter, and forums. EMR infectious conjunctivitis had an earlier circular mean week for younger age groups than for older, but this variation was not observed as much for younger compared to older ages in Twitter and forums (see Table 5 and Fig. 5). Additionally, the mean week of infectious and allergic

TABLE 3. EMR Seasonal Characteristics of Infectious, Allergic, and Other Conjunctivitis Groups, as Well as Influenza, by Age Group

Condition	Age	Mean	Circ. Var.	Amplitude	*P* Value
Allergic	0–5	July 2 (May 9–August 22)	0.85 (0.75–0.94	0.78 (0.59–0.91)	0.052
Allergic	6–17	June 16 (May 31–July 2)	0.81 (0.77–0.85)	0.69 (0.61–0.78)	<0.001
Allergic	18–39	June 4 (May 15–June 25)	0.86 (0.8–0.91)	0.60 (0.47–0.72)	<0.001
Allergic	40+	May 13 (April 29–May 27)	0.86 (0.82–0.89)	0.49 (0.42–0.59)	<0.001
Infectious	0–5	March 9 (March 5–March 15)	0.64 (0.62–0.68)	0.79 (0.75–0.84)	<0.001
Infectious	6–17	March 18 (March 10–March 26)	0.74 (0.71–0.78)	0.70 (0.64–0.77)	<0.001
Infectious	18–39	March 30 (March 10–April 22)	0.89 (0.85–0.93)	0.47 (0.37–0.6)	<0.001
Infectious	40+	March 29 (March 16–April 10)	0.87 (0.84–0.9)	0.47 (0.40–0.57)	<0.001
Other	0–5	January 4 (January 5–December 28)	0.91 (0.77–0.98)	0.87 (0.72–0.97)	0.029
Other	6–17	June 8 (April 27–July 23)	0.69 (0.55–0.86)	0.82 (0.71–0.97)	0.24
Other	18–39	June 20 (May 3–July 31)	0.85 (0.74–0.94)	0.64 (0.56–0.85)	0.013
Other	40+	June 24 (May 22–July 31)	0.86 (0.78–0.93)	0.54 (0.44–0.75)	0.02
Flu	0–5	February 3 (January 26–February 9)	0.29 (0.24–0.38)	0.98 (0.96–1)	<0.001
Flu	6–17	February 4 (January 30–February 9)	0.23 (0.19–0.3)	0.98 (0.98–0.99)	<0.001
Flu	18–39	January 30 (January 24–February 5)	0.35 (0.29–0.4)	0.98 (0.96–0.99)	<0.001
Flu	40+	January 31 (January 26–February 5)	0.36 (0.32–0.41)	0.96 (0.94–0.98)	<0.001

95% confidence intervals for the estimates are given in table; *P* values indicate strength of evidence for seasonality.

conjunctivitis consistently occurred slightly later for EMR than for Twitter and forums (see Tables 2, 5).

We compared seasonal features of infectious conjunctivitis to influenza, which is known to be highly seasonal, as well as to other diseases considered less seasonal. Age-based variation of circular mean week, circular variance and relative amplitude observed for infectious conjunctivitis were less varied for influenza and more pronounced (see Fig. 4D, Table 3). Comparing circular variance and relative amplitude, these infectious conjunctivitis seasonal features were lower than for influenza, but higher than for allergic conjunctivitis, other conjunctivitis, and nonconjunctivitis eye disease (see Tables 2, 3; Fig. 4E).

DISCUSSION

Social media posts from Twitter or forums, when classified as pertaining to infectious conjunctivitis, had similar seasonal characteristics (mean week, circular variance, and amplitude) to each other, and to seasonal infectious conjunctivitis occurrence. In the same way, the occurrence of social media

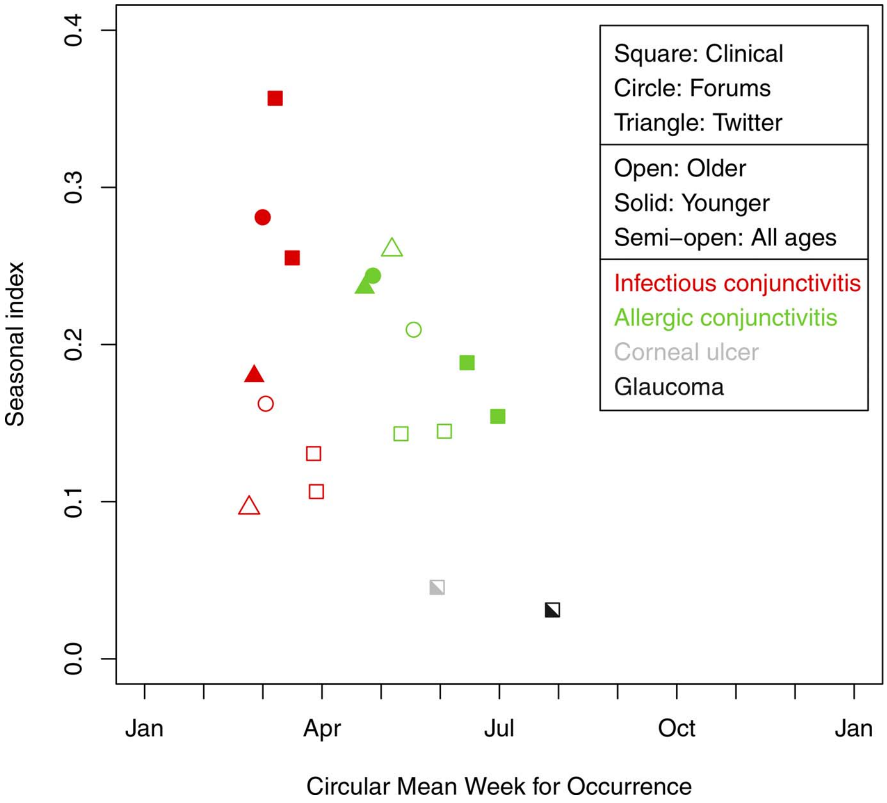

FIGURE 5. Timing and degree of seasonality for selected clinical and social media data. The circular mean week of occurrence is shown on the *horizontal axis*; the *vertical axis* displays a measure of degree of seasonality (one minus the circular variance; higher location indicates greater seasonality). Infectious conjunctivitis is shown in *red*, allergic conjunctivitis in *green*. Glaucoma and corneal ulcers are shown for comparison.

Table 4. Spearman Rank Correlation of Detrended Residuals for Clinical Counts, Twitter, and Forum Posts, for Infectious and Allergic Conjunctivitis

	Infectious Conjunctivitis			**Allergic Conjunctivitis**		
	Clinical	**Twitter**	**Forums**	**Clinical**	**Twitter**	**Forums**
Infectious						
Clinical	–	0.64	0.4	0.21	0.35	0.33
Twitter	(0.44, 0.75)	–	0.48	0.033	0.17	0.17
Forums	(0.23, 0.55)	(0.29, 0.62)	–	−0.029	−0.024	0.092
Allergic						
Clinical	(0.014, 0.39)	(−0.17, 0.23)	(−0.19, 0.13)	–	0.55	0.5
Twitter	(0.11, 0.54)	(−0.092, 0.41)	(−0.24, 0.2)	(0.39, 0.66)	–	0.69
Forums	(0.12, 0.51)	(−0.064, 0.38)	(−0.16, 0.36)	(0.33, 0.61)	(0.52, 0.78)	–

Single numerals represent the estimated Spearman rank correlations, two numbers in parentheses represent the corresponding 95% confidence intervals.

posts classified as related to allergic conjunctivitis showed similar timing to the occurrence of clinical cases of allergic conjunctivitis. The mean week of occurrence of infectious conjunctivitis consistently occurred approximately 10 weeks before allergic conjunctivitis, for any data source (clinical, Twitter, or Forum). Broken down by age, social media posts also showed similar seasonal characteristics to corresponding clinical age groups. Our results suggest that social media data regarding conjunctivitis may mirror underlying clinical phenomena.

Despite finding seasonal similarities of posts and clinical conjunctivitis data, some differences existed. For example, we found a smaller ratio of infectious to allergic conjunctivitis for posts than for clinical data. We note that clinical data includes the true calendar age, whereas social media analysis may involve inferred ages. Other biases may exist: perhaps individuals posting concerning allergic conjunctivitis seek health care with a lower frequency than those posting with infectious conjunctivitis. Clinically, seasonality of infectious conjunctivitis for younger ages (0–5, pediatrics) showed an earlier typical occurrence, and exhibited a more pronounced seasonality (smaller circular variance) than older ages (6–40+, ophthalmology).

Several limitations apply to our findings. Boolean and machine-learned classification of posts into disease and age groups introduces unavoidable misclassification. Our human-derived validations of posts indicated that misclassification appears to be uncommon for Twitter posts. They also indicated that for both Twitter and forum posts, whenever humans agreed that a post was either about infectious or allergic conjunctivitis, the machine almost always agreed. However, as indicated in the “Results” section, for a substantial fraction of those forum posts (but not of Twitter posts) that were machine coded as conjunctivitis, the humans classified these as uncertain or not about conjunctivitis. Despite this, forum data still appear to support the distinct seasonal infectious and allergic conjunctivitis relationships seen in Twitter and clinical data. Future study could further refine the forum queries, increasing the agreement rate for forum posts and eliminating nonspecific posts that may have no distinct seasonal pattern. Moreover, posts, as well as our EMR clinical data, may represent limited portions of the national population.[34–37]

Regarding clinical data, we have compared populations and seasonal patterns of acute conjunctivitis cases for emergency medicine within our dataset to that of the national emergency department sample dataset, and found similar populations by

Table 5. Seasonal Characteristics of Conjunctivitis in Clinical EMR and Social Media Posts, by Age

Data Source	**Age Group**	**Mean Week**	**Circular Variance**	**Relative Amplitude**	***P* Value**
Infectious conjunctivitis by age					
Clinical	0–5	March 9 (March 5–15)	0.64 (0.62–0.68)	0.79 (0.75–0.84)	<0.001
Twitter	Younger	February 26 (February 19–March 6)	0.82 (0.80–0.84)	0.57 (0.52–0.62)	<0.001
Forums	Younger	March 3 (February 19–March 14)	0.72 (0.66–0.77)	0.78 (0.69–0.84)	<0.001
Clinical	6–17	March 18 (March 10–26)	0.74 (0.71–0.78)	0.70 (0.64–0.77)	<0.001
Clinical	18–39	March 30 (March 10–April 22)	0.89 (0.85–0.93)	0.47 (0.37–0.60)	<0.001
Clinical	40+	March 29 (March 16–April 10)	0.87 (0.84–0.90)	0.47 (0.40–0.57)	<0.001
Twitter	Older	February 24 (February 15–March 2)	0.90 (0.89–0.92)	0.37 (0.34–0.41)	<0.001
Forums	Older	March 4 (February 21–March 17)	0.84 (0.80–0.87)	0.56 (0.47–0.65)	<0.001
Allergic conjunctivitis by age					
Clinical	0–5	July 2 (May 9–August 22)	0.85 (0.75–0.94)	0.78 (0.59–0.91)	0.052
Twitter	Younger	April 24 (April 16–May 4)	0.76 (0.72–0.80)	0.71 (0.66–0.77)	<0.001
Forums	Younger	April 29 (April 13–May 14)	0.76 (0.69–0.82)	0.71 (0.60–0.82)	<0.001
Clinical	6–17	June 16 (May 31–July 2)	0.81 (0.77–0.85)	0.69 (0.61–0.78)	<0.001
Clinical	18–39	June 4 (May 15–June 25)	0.86 (0.80–0.91)	0.60 (0.47–0.72)	<0.001
Clinical	40+	May 13 (April 29–May 27)	0.86 (0.82–0.89)	0.49 (0.42–0.59)	<0.001
Twitter	Older	May 8 (May 6–May 11)	0.74 (0.73–0.75)	0.73 (0.71–0.75)	<0.001
Forums	Older	May 20 (May 10–May 30)	0.79 (0.75–0.83)	0.60 (0.55–0.69)	<0.001

As in the text, the circular mean, circular variance, and (relative) amplitude are derived from estimated cyclic splines; the *P* value reflects the test of the null hypothesis of nonseasonality.

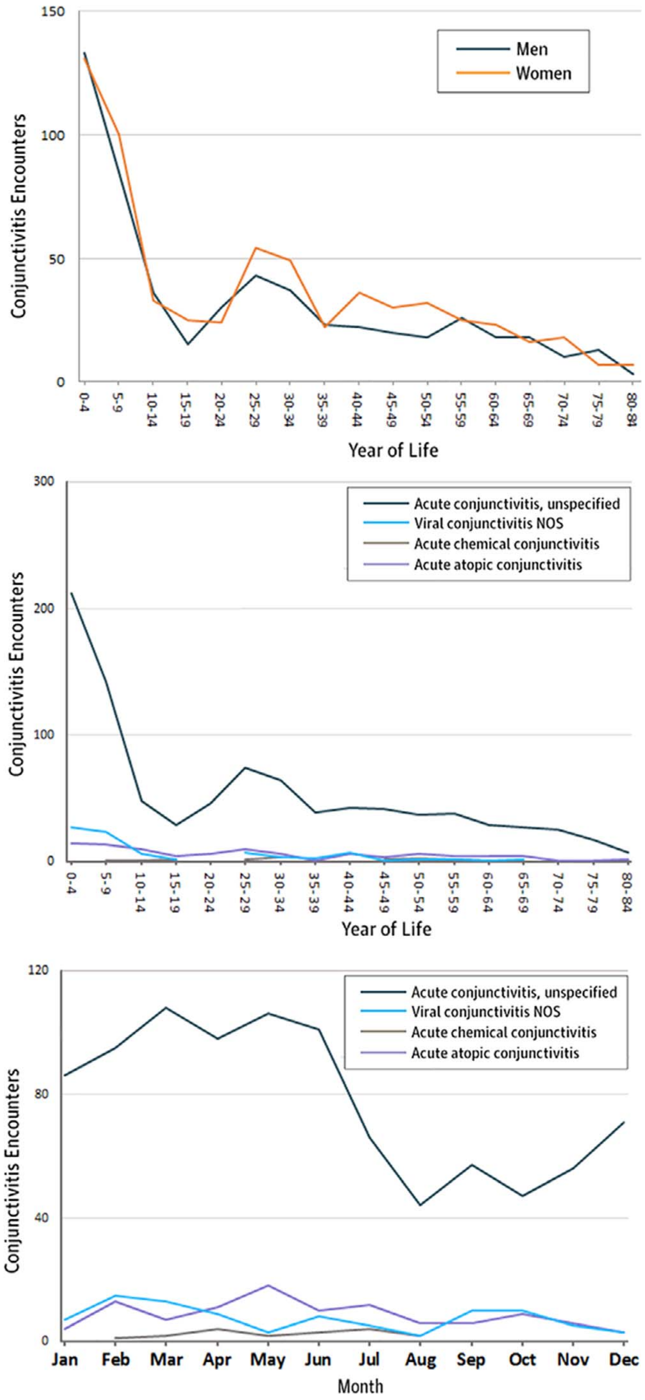

FIGURE 6. Age-based and monthly UCSF EMR emergency medicine data, for comparison to national clinical emergency department data (see Ref. 32). *Top*: age distributions stratified by sex (for comparison, see figure 1a of Ref. 32). *Center*: age distributions stratified by conjunctivitis diagnosis code groups (for comparison, see figure 1b of Ref. 32). *Bottom*: monthly distributions stratified by conjunctivitis diagnosis code groups (for comparison, see figure 2b of Ref. 32).

age and by conjunctivitis group, suggesting that (at least for emergency medicine) the clinical data used in this current study may at least be partially representative of national clinical data (see Fig. 6; figures 1a, 1b, 2b in Ramirez et al.[32]). Additionally, clinical data also could include diagnoses from multiple specialties, whose providers may exhibit differences in diagnosis and treatment of the same condition.[38] For any given episode, the patient's self-diagnosis (as reflected in the subject of the tweet or post) may differ from the clinical diagnosis, and it is possible that patients are more likely to consider an episode "infectious" than allergic. A similar phenomenon may explain the relative overprescribing of antibiotics even by noneye care specialists versus eye specialists in the treatment of conjunctivitis.[38] Future studies could consider use of national clinical data representing all specialties as well as other social media, search, and weblog data. Alternatively, a future study comparing specific geographic or demographic sectors, from diverse clinical or social media platforms, could identify important differences in occurrence to potentially guide more targeted eye health care or policy implementations.

Despite these and other limitations, the findings of our study suggest future use of machine learning and refined Boolean query could allow for even more granular understanding of prevalence and seasonal patterns of different conjunctivitis etiologies. This, in turn, may greatly enhance identification of infectious conjunctivitis occurrence outside normal seasonal patterns for age or geographic subgroups, potentially allowing for improved outbreak detection by combined monitoring and analysis of clinical and Internet-based data.

Acknowledgments

The authors thank Suling Wang for assistance.

Supported in part by Grant 1R01EY024608-01A1 from the National Institutes of Health National Eye Institute (NIH-NEI); Grant EY002162 (Core Grant for Vision Research) from the NIH-NEI; and an unrestricted grant from Research to Prevent Blindness, through the University of California San Francisco Information Technology Enterprise Information Analytics Departments Research Data Browser and Clinical Data Research Consultation Services.

Disclosure: **M.S. Deiner**, None; **S.D. McLeod**, None; **J. Chodosh**, None; **C.E. Oldenburg**, None; **C.A. Fathy**, None; **T.M. Lietman**, None; **T.C. Porco**, None

References

1. Brownstein JS, Freifeld CC. HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveill*. 2007;12:E071129.5.
2. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the web for public health surveillance. *N Engl J Med*. 2009;360:2153–2155.
3. Madan A, Cebrian M, Lazer D, Pentland A. Social sensing for epidemiological behavior change. *Proc ACM Int Conf Ubiquitous Comput*. 2010:291–300.
4. Khan K, McNabb SJN, Memish ZA, et al. Infectious disease surveillance and modelling across geographic frontiers and scientific specialties. *Lancet Infect Dis*. 2012;12:222–230.
5. Sadilek A, Kautz H, Bigham JP. Modeling the interplay of people's location, interactions, and social ties. In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. Beijing, China: AAAI Press; 2013:3067–3071.
6. Hartley DM, Nelson NP, Arthur RR, et al. An overview of internet biosurveillance. *Clin Microbiol Infect*. 2013;19:1006–1013.
7. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and Internet-based data in global systems for public health surveillance: a systematic review. *Milbank Q*. 2014;92:7–33.

8. Nuti SV, Wayda B, Ranasinghe I, et al. The use of Google trends in health care research: A systematic review. *PLoS One*. 2014;9:e109583.
9. Brownstein JS, Mandl KD. Reengineering real time outbreak detection systems for influenza epidemic monitoring. *AMIA Annu Symp Proc*. 2006: 866.
10. Brownstein JS, Freifeld CC, Madoff LC. Influenza A (H1N1) virus, 2009—online monitoring. *N Engl J Med*. 2009;360: 2156.
11. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457:1012–1014.
12. Barboza P, Vaillant L, Le Strat Y, et al. Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PLoS One*. 2014;9:e90536.
13. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. *PLoS Comput Biol*. 2014;10:e1003892.
14. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol*. 2015;11:e1004513.
15. Hoen AG, Keller M, Verma AD, Buckeridge DL, Brownstein JS. Electronic event-based surveillance for monitoring dengue, Latin America. *Emerg Infect Dis*. 2012;18:1147–1150.
16. HealthMap. Available at: http://www.healthmap.org/site/about. Accessed July 19, 2017.
17. Leffler CT, Davenport B, Chan D. Frequency and seasonal variation of ophthalmology-related Internet searches. *Can J Ophthalmol*. 2010;45:274–279.
18. Kang MG, Song WJ, Choi S, et al. Google unveils a glimpse of allergic rhinitis in the real world. *Allergy*. 2015;70:124–128.
19. McGregor F, Somner JEA, Bourne RR, Munn-Giddings C, Shah P, Cross V. Social media use by patients with glaucoma: What can we learn? *Ophthalmic Physiol Opt*. 2014;34:46–52.
20. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance tools emerging from search engines and social media data for determining eye disease patterns. *JAMA Ophthalmology*. 2016;134:1024–1030.
21. Bousquet J, Caimmi DP, Bedbrook A, et al. Pilot study of mobile phone technology in allergic rhinitis in European countries: the MASK-rhinitis study. *Allergy*. 2017;72:857–865.
22. Benke KK. Uncertainties in big data when using Internet surveillance tools and social media for determining patterns in disease incidence. *JAMA Ophthalmol*. 2017;135:402.
23. Sommer A. The utility of "big data" and social media for anticipating, preventing, and treating disease. *JAMA Ophthalmol*. 2016;134:1030–1031.
24. Deiner MS, Lietman TM, Porco TC. Uncertainties in big data when using Internet surveillance tools and social media for determining patterns in disease incidence-reply. *JAMA Ophthalmology*. 2017;135:402–403.
25. Smith AF, Waycaster C. Estimate of the direct and indirect annual cost of bacterial conjunctivitis in the United States. *BMC Ophthalmol*. 2009;9:13.
26. Benzekri R, Belfort R Jr, Ventura CV, et al. Manifestations oculaires du virus Zika: Où en sommes-nous? *J Fr Ophtalmol*. 2017;40:128–145.
27. Crimson Hexagon. Available at: http://www.crimsonhexagon.com. Accessed July 19, 2017.
28. Hopkins D, King G. A method of automated nonparametric content analysis for social science. *Am J Pol Sci*. 2010;54:229–247.
29. Firat A, Brooks M, Bingham C, Herdagdelen A, King G. Systems and methods for calculating category proportions. 2014.
30. Hilbe JM. *Negative Binomial Regression*. Cambridge: Cambridge University Press; 2011.
31. Wood SN. *Generalized Additive Models: An Introduction with R, Second Edition*. Boca Raton, FL: CRC Press; 2017.
32. Ramirez DA, Porco TC, Lietman TM, Keenan JD. Epidemiology of conjunctivitis in US emergency departments. *JAMA Ophthalmol*. 2017;135:1119–1121.
33. Crimson Hexagon Topic Wheel. Available at: https://help.crimsonhexagon.com/hc/en-us/articles/203641365-Explore-Tab-Topic-Wheel-Section. Accessed July 19, 2017.
34. Sadah SA, Shahbazi M, Wiley MT, Hristidis V. Demographic-based content analysis of web-based health-related social media. *J Med Internet Res*. 2016;18:e148.
35. Pew Research. Social media fact sheet. Available at: http://www.pewinternet.org/fact-sheet/social-media/. Accessed July 19, 2017.
36. Sadah SA, Shahbazi M, Wiley MT, Hristidis V. A study of the demographics of web-based health-related social media users. *J Med Internet Res*. 2015;17:e194.
37. Pew Research. Social media update 2016. Available at: http://www.pewinternet.org/2016/11/11/social-media-update-2016/. Accessed July 19, 2017.
38. Shekhawat NS, Shtein RM, Blachley TS, Stein JD. Antibiotic prescription fills for acute conjunctivitis among enrollees in a large United States managed care network. *Ophthalmology*. 2017;124:1099–1107.
39. Fisher NI. *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press; 1993.
40. Leith CE. The standard error of time-average estimates of climatic means. *J Appl Meteorol*. 1973;12:1066–1069.
41. Crimson Hexagon Location Methodology. Available at: https://help.crimsonhexagon.com/hc/en-us/articles/203952525-Location-Methodology. Accessed July 19, 2017.
42. Crimson Hexagon Language Methodology. Available at: https://help.crimsonhexagon.com/hc/en-us/articles/202772699-Language-Filter-How-Does-it-Work. Accessed July 19, 2017.

Appendix

Statistical Calculation

For observation i, denote the observed count (clinical, Twitter, or forums) by Y_i and the occurrence time as t_i, with corresponding phase angle ϕ_i (with 0 radians corresponding to January 1). Negative binomial regression with both temporal trend and seasonal spline basis functions then yields the detrended, smoothed estimate of the expected count at phase ϕ as $\hat{f}(\phi) \equiv \sum_{j}^{k} \hat{a}_j s_j(\phi)$, where $\hat{a}_j$ is the estimated regression coefficient for the jth spline basis function ($k = 8$ is the number of spline basis functions used). Corresponding to $\hat{f}(\phi)$ is the estimated first circular moment,[39] a complex valued quantity given by $\hat{m}_1 \equiv \int_0^{2\pi} (\cos(\phi) + i\sin(\phi))\hat{f}(\phi)\, d\phi$ (note: $i = \sqrt{-1}$). The estimated circular mean phase $\hat{\phi}$ was computed from the first circular moment $\hat{m}_1$ according to $\hat{\phi} = \arg(\hat{m}_1)$, and reported in days or weeks (instead of degrees or radians). The circular variance v is defined as $1 - |m_1|$, and measures the lack of seasonality. We defined relative amplitude as simply $|\max_\phi f(\phi) - \min_\phi f(\phi)|/|\max_\phi(\phi)|$ (i.e., the ratio of the peak-to-trough distance to the total peak). Residuals were examined for autocorrelation. Standard errors for overall seasonal summary statistics (circular mean time of occurrence, circular variance, and relative amplitude) were computed using Monte Carlo simulation based on the estimated standard errors for the regression coefficients for the spline basis functions which yield the fitted seasonal curve (from which the summary statistics were derived). Standard errors for these quantities were inflated based on an autocorrelation-based effective sample size formula.[40]

Validation of Machine-Coded Classifications

Two human raters reviewed materials from the American Academy of Ophthalmology website concerning infectious and allergic conjunctivitis, followed by two training sessions using randomly chosen conjunctivitis posts. We conducted a modified Delphi session as follows in which each rater classified each post from a common set of randomly chosen posts as allergic (A); infectious (I); or unsure/other (O). Raters were masked to the machine-coded classifications and to other raters. After all ratings were completed, a moderator identified posts for which disagreement occurred. The moderator was masked to the machine-coded classifications. For each post on which the raters disagreed, the moderator elicited follow-up comments (one to two sentences) from raters in a random order, followed by an opportunity for the raters to update their classifications. A final round was conducted in the same way, after which the data set was locked. From this final human rated dataset, the human consensus classification $C(x,y)$ for two ratings x and y was defined by $C(x,x)=x$, $C(x,O)=x$, and $C(A,I)=C(I,A)=O$. The sample size for tweets was fixed in advance at 128, which provides a confidence interval half-width of under 0.1 for a proportion of 0.5; half this number of forum posts were scored (since the forum posts are, on average, much longer and take more time to assess).

Social Media Geographic Location

Using Crimson Hexagon's geocoding and language algorithms,[41,42] we sought to maximize USA geographic specificity in our query. Twitter results were filtered to include only tweets which contained "Locations: United States of America". Blogs, forums, and comments results, for which reliable geocoding was not available, were filtered to include just those posts which contained "Language: English" (please see Supplementary Material for additional query details).