JAMA Ophthalmology | Original Investigation

Surveillance Tools Emerging From Search Engines and Social Media Data for Determining Eye Disease Patterns

Michael S. Deiner, PhD; Thomas M. Lietman, MD; Stephen D. McLeod, MD; James Chodosh, MD, MPH; Travis C. Porco, PhD, MPH

IMPORTANCE Internet-based search engine and social media data may provide a novel complementary source for better understanding the epidemiologic factors of infectious eye diseases, which could better inform eye health care and disease prevention.

OBJECTIVE To assess whether data from internet-based social media and search engines are associated with objective clinic-based diagnoses of conjunctivitis.

DESIGN, SETTING, AND PARTICIPANTS Data from encounters of 4143 patients diagnosed with conjunctivitis from June 3, 2012, to April 26, 2014, at the University of California San Francisco (UCSF) Medical Center, were analyzed using Spearman rank correlation of each weekly observation to compare demographics and seasonality of nonallergic conjunctivitis with allergic conjunctivitis. Data for patient encounters with diagnoses for glaucoma and influenza were also obtained for the same period and compared with conjunctivitis. Temporal patterns of Twitter and Google web search data, geolocated to the United States and associated with these clinical diagnoses, were compared with the clinical encounters. The a priori hypothesis was that weekly internet-based searches and social media posts about conjunctivitis may reflect the true weekly clinical occurrence of conjunctivitis.

MAIN OUTCOMES AND MEASURES Weekly total clinical diagnoses at UCSF of nonallergic conjunctivitis, allergic conjunctivitis, glaucoma, and influenza were compared using Spearman rank correlation with equivalent weekly data on Tweets related to disease or disease-related keyword searches obtained from Google Trends.

RESULTS Seasonality of clinical diagnoses of nonallergic conjunctivitis among the 4143 patients (2364 females [57.1%] and 1776 males [42.9%]) with 5816 conjunctivitis encounters at UCSF correlated strongly with results of Google searches in the United States for the term *pink eye* (ρ, 0.68 [95% CI, 0.52 to 0.78]; $P < .001$) and correlated moderately with Twitter results about *pink eye* (ρ, 0.38 [95% CI, 0.16 to 0.56]; $P < .001$) and with clinical diagnosis of influenza (ρ, 0.33 [95% CI, 0.12 to 0.49]; $P < .001$), but did not significantly correlate with seasonality of clinical diagnoses of allergic conjunctivitis diagnosis at UCSF (ρ, 0.21 [95% CI, −0.02 to 0.42]; $P = .06$) or with results of Google searches in the United States for the term *eye allergy* (ρ, 0.13 [95% CI, −0.06 to 0.32]; $P = .19$). Seasonality of clinical diagnoses of allergic conjunctivitis at UCSF correlated strongly with results of Google searches in the United States for the term *eye allergy* (ρ, 0.44 [95% CI, 0.24 to 0.60]; $P < .001$) and *eye drops* (ρ, 0.47 [95% CI, 0.27 to 0.62]; $P < .001$).

CONCLUSIONS AND RELEVANCE Internet-based search engine and social media data may reflect the occurrence of clinically diagnosed conjunctivitis, suggesting that these data sources can be leveraged to better understand the epidemiologic factors of conjunctivitis.

Invited Commentary page 1030

Supplemental content at jamaophthalmology.com

Author Affiliations: Department of Ophthalmology, University of California San Francisco (Deiner, Lietman, McLeod, Porco); F. I. Proctor Foundation, University of California San Francisco (Lietman, McLeod, Porco); Department of Epidemiology and Biostatistics, University of California San Francisco (Lietman, Porco); Global Health Sciences, University of California San Francisco (Lietman); Massachusetts Eye and Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston (Chodosh).

Corresponding Author: Travis C. Porco, PhD, MPH, F.I. Proctor Foundation, University of California San Francisco, 513 Parnassus, Medical Sciences Room S334, San Francisco, CA 94143 (travis.porco@ucsf.edu).

JAMA Ophthalmol. 2016;134(9):1024-1030. doi:10.1001/jamaophthalmol.2016.2267
Published online July 14, 2016.

Conjunctivitis is one of the most common eye diseases in the United States. It often causes eye pain, discomfort, and temporary vision impairment, and rarely causes permanent conjunctival and corneal scarring. Conjunctivitis has 3 predominant forms: bacterial,[1] viral,[2] and allergic,[3-5] each with unique characteristics.[6] It contributes substantial annual costs in the United States affecting health care, the workforce, and education.[7-9] Children with conjunctivitis are prohibited from attending school, causing parents to miss work or pay for childcare; a study from 2014 calculated that conjunctivitis has an overall annual US medical cost of $800 million and causes 3.5 million missed school days and 8.5 million missed work days annually, with estimated annual lost wages of $1.9 billion.[9] Epidemics of severe conjunctivitis appear to be on the rise in some countries and are endemic in others.[2] Despite the effect of conjunctivitis in the United States, to our knowledge, no primary eye-specific infectious diseases, including conjunctivitis, are regularly tracked by the Centers for Disease Control and Prevention. Therefore, seasonality, incidence, and frequency of conjunctivitis epidemics in the United States are not well documented. However, the prevalence of conjunctivitis in the United States is estimated at approximately 2.2%, and conjunctivitis is estimated to account for approximately 1% of all emergency department and primary care physician visits.[9]

In the past decade, public health and research organizations have begun to supplement standard health and disease monitoring and epidemiologic reporting with complementary information obtained through digital surveillance of social media, including both passive (geocoded web traffic and keywords from social networks and search engines, news feeds, and blogs) and active (participatory electronic surveys and electronic medical record [EMR] registries) sources.[10-17] These approaches have the potential to improve understanding of epidemiologic factors and to detect outbreaks much sooner than the traditional criterion standard methods, and have reportedly been shown to detect and predict influenza and other outbreaks weeks in advance of traditional Centers for Disease Control and Prevention methods, complementing traditional reporting.[18-23] However, it is for traditionally less well-monitored or less-reported infectious diseases where social media data may have the greatest potential to provide epidemiologic information,[24] as is currently the case in the United States for infectious eye diseases.

Some studies have begun investigating the role of social media as related to ophthalmology[25] and some nonprofit and commercial systems are tracking conjunctivitis geospatially, for example, by using news feeds and randomly submitted reports.[26] Previous analyses of data for eye-related terms from Google Trends before 2009 have established the seasonality of conjunctivitis, with a peak in colder weather.[27] However, it has not been confirmed if this pattern continued in subsequent years, or how it corresponds to the seasonal patterns of incidence as seen in clinical practice. It is also unknown if other sources of social media, such as Twitter, may add to our understanding. Other studies have investigated allergic rhinitis, which has an ocular component, and have suggested a strong correlation of allergic rhinitis with internet-based Google searches, web traffic logs, and other related terms such as *medications*.[28] We compare the seasonality of conjunctivitis in online searches in the United States with the seasonality observed in an EMR system from a tertiary care center. We compare that correlation of conjunctivitis-related seasonality with that of other eye-related and of non–eye-related online searches and EMR data as well as with Tweets about pink eye. We also perform a subanalysis of allergic conjunctivitis and nonallergic conjunctivitis.

Key Points

Question Can internet-based data from social media and search engines provide novel sources of epidemiologic factors of infectious eye diseases, associated with objective clinic-based diagnoses?

Findings Seasonality of clinical diagnoses of nonallergic conjunctivitis from electronic medical records correlated strongly with results of Google searches in the United States for the term *pink eye*, and correlated moderately with Tweets about pink eye and with clinical diagnosis of influenza. Seasonality of clinical diagnoses of allergic conjunctivitis from electronic medical records correlated strongly with results of Google searches in the United States for the term *eye allergy*.

Meaning Internet-based data from search engines and social media may provide a novel complementary source for understanding the epidemiologic factors of infectious eye disease.

Methods

Data Acquisition

EMR Clinical Data

With approval from the University of California San Francisco (UCSF) Institutional Review Board, we obtained total weekly counts of all encounters with diagnosis names containing the string "conjunctivi" for June 3, 2012, to April 26, 2014, from the UCSF EMR. Resulting encounters were grouped into allergic and nonallergic conjunctivitis, based on whether the conjunctivitis diagnosis name contained the string "allerg." In addition to conjunctivitis diagnosis name encounters, we also obtained total weekly counts for the same period from the UCSF EMR for glaucoma and for influenza. Informed consent was waived because patient data were deidentified.

Google Search Data

Results of searches were obtained from Google Trends[17] using the United States as the search location for the same period as the EMR data. The keywords used were *pink eye*, *eye allergy*, *flu*, and *eye drops*. We also searched Google Trends using Australia as the location for the keyword *conjunctivitis* during the same period (*pink eye* is rarely used to describe conjunctivitis in Australia).

Twitter

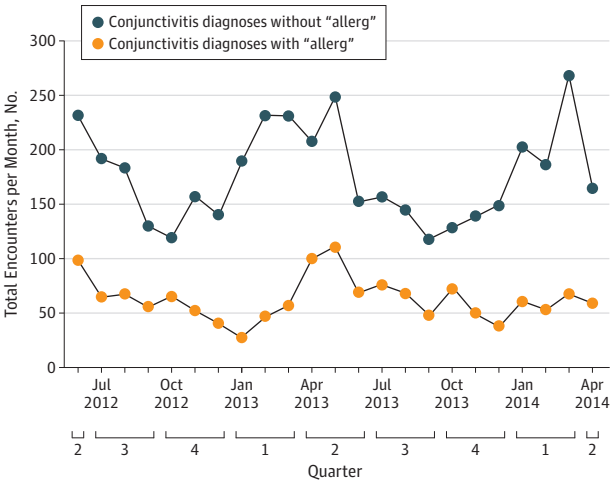
A random sample of 6441 Tweets, geolocated for the United States and enriched via the Boolean query to include Tweets with first-person statements regarding having or getting conjunctivitis (and enriched to exclude Tweets regarding celebrities, cinematic topics from popular culture, animals, reposts of

Table 1. Demographics of Patients With Clinical Conjunctivitis

Characteristic	No. (%) Nonallergic Conjunctivitis (n = 3131)	Allergic Conjunctivitis (n = 1009)	Total (N = 4140)[a]
Sex			
Female	1716 (54.8)	648 (64.2)	2364 (57.1)
Male	1415 (45.2)	361 (35.8)	1776 (42.9)
Age group, y			
<5	770 (24.6)	22 (2.2)	792 (19.1)
5-19	622 (19.9)	240 (23.8)	862 (20.8)
20-34	422 (13.5)	115 (11.4)	537 (13.0)
35-49	501 (16.0)	169 (16.7)	670 (16.2)
50-64	406 (13.0)	212 (21.0)	618 (14.9)
65-79	291 (9.3)	166 (16.5)	457 (11.0)
80-94	112 (3.6)	80 (7.9)	192 (4.6)
95-109	9 (0.3)	5 (0.5)	14 (0.3)
Race/ethnicity			
Native American, Alaska Native	12 (0.4)	2 (0.2)	14 (0.4)
Asian	500 (17.1)	266 (28.4)	766 (19.8)
Black or African American	308 (10.5)	96 (10.3)	404 (10.5)
Native Hawaiian, Pacific Islander	47 (1.6)	24 (2.6)	71 (1.8)
Other	615 (21.0)	195 (20.8)	810 (21.0)
White	1441 (49.3)	353 (37.7)	1794 (46.5)

[a] Data for June 3, 2012, to April 26, 2014, based on 4143 unique patients with 5816 encounters of diagnosed conjunctivitis. Numbers do not total 4143 owing to a small number of patients with missing records for particular fields.

Figure 1. Number of Diagnoses of Nonallergic and Allergic Conjunctivitis



Number of diagnoses of nonallergic and allergic conjunctivitis in the University of California San Francisco electronic medical record, June 3, 2012, to April 26, 2014, based on all 5816 diagnoses (data for April 2014 end on the 26th; the total for full month would likely be higher than shown). Diagnoses of nonallergic conjunctivitis were those without the string "allerg" in the electronic medical record; diagnoses of allergic conjunctivitis were those with the string "allerg" in the electronic medical record.

URLs, and retweets) was obtained through the Crimson Hexagon platform[29] for the same period as the EMR data (see the eAppendix in the Supplement for the detailed query). Clinical EMR data from UCSF were obtained in the fall of 2014, while data from Twitter and Google searches were obtained in the spring of 2015.

Statistical Analysis

We conducted Spearman rank correlation of each weekly observation. Time series bootstrap (with a fixed window of 2) was conducted to construct 95% CIs and *P* values.[30] All computations were conducted in *R*, version 3.1 for MacIntosh (*R* Foundation for Statistical Computing).

Results

Demographics of Clinical Diagnoses

Demographic characteristics of the UCSF group are summarized in **Table 1**. The UCSF conjunctivitis query resulted in a data set containing 4143 patients with 5816 conjunctivitis encounters. Patients were from 67 departments, including general pediatrics (1106 [26.7%]), ophthalmology (840 [20.3%]), and general internal medicine (693 [16.7%]). The cohort comprised 2364 females (57.1%), 1776 males (42.9%), 1794 white patients (46.5%), 766 Asian patients (19.8%), 404 black or African American patients (10.5%), 14 Native American or Alaska Native patients (0.4%), and 71 Native Hawaiian or Pacific Islander patients (1.8%); 810 patients (21.0%) had no information on race/ethnicity. We found evidence of a significant age difference between patients with nonallergic and allergic conjunctivitis (*P* < .001, Wilcoxon rank-sum test). The mean age of patients with nonallergic conjunctivitis was 30.3 years (95% CI, 29.4-31.2), and of allergic patients, 43.6 years (95% CI, 42.0-45.2). Table 1 shows the largest difference between patients with allergic and nonallergic conjunctivitis at the youngest age class (<5 years). Patients with allergic and nonallergic conjunctivitis also differed by race/ethnicity (*P* < .001, Fisher exact test); for example, among the group with allergic conjunctivitis, Asian race made up a higher percentage (266 [28.4%]) than in the group with nonallergic conjunctivitis (500 [17.1%]), while, inversely, white patients made up a higher percentage of the group with nonallergic conjunctivitis (1441 [49.3%]) than of the group with allergic conjunctivitis (353 [37.7%]). In addition, we found evidence of a difference by sex (*P* < .001, Fisher exact test), with 648 females (64.2%) in the group with allergic conjunctivitis vs 1716 females (54.8%) in the group with nonallergic conjunctivitis. The seasonality of all UCSF encounters of patients with allergic conjunctivitis and those with nonallergic conjunctivitis is shown in **Figure 1**. The frequency of encounters with patients with nonallergic conjunctivitis fluctuated over time, roughly doubling in size from fall to spring for each year observed and then returning to fall levels. Cases of allergic conjunctivitis followed a similar pattern (increase from fall to spring, then back to fall levels), but the seasonality appeared delayed behind the encounters with patients with nonallergic conjunctivitis by approximately 2 months, and perhaps with more varied levels of fluctuation between years.

Spearman Rank Correlation Comparison

Table 2 compares the clinical diagnoses with results of Google searches and Tweets. For patient encounters at UCSF with nonallergic conjunctivitis, we found the strongest correlations with

Table 2. Spearman Rank Correlation (95% CI) for Weekly UCSF Encounters With Eye Disease Diagnoses vs Disease-Related Google Searches and Tweets[a]

Characteristic	Nonallergic Conjunctivitis Diagnoses	Allergic Conjunctivitis Diagnoses	Glaucoma Diagnoses	Influenza Diagnoses	Google USA Search for *Pink Eye*	Google Australia Search for *Conjunctivitis*	Google USA Search for *Eye Allergy*	Google USA Search for *Flu*	Google USA Search for *Eye Drops*	Twitter USA Posts About *Pink Eye*
Nonallergic conjunctivitis diagnoses	1	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
Allergic conjunctivitis diagnoses	0.21 (−0.02 to 0.42)	1	[b]	[b]	[b]	[b]	[b]	[b]	[b]	[b]
Glaucoma diagnoses	−0.02 (−0.24 to 0.22)	0.24 (0.02 to 0.44)	1	[b]	[b]	[b]	[b]	[b]	[b]	[b]
Influenza diagnoses	0.33 (0.12 to 0.49) [b]	−0.30 (−0.49 to −0.08)[c]	−0.04 (−0.25 to 0.18)	1	[b]	[b]	[b]	[b]	[b]	[b]
Google USA search for *pink eye*	0.68 (0.52 to 0.78)[c]	0 (−0.25 to 0.23)	−0.32 (−0.51 to −0.09)[c]	0.47 (0.25 to 0.66)[c]	1	[b]	[b]	[b]	[b]	[b]
Google Australia search for *conjunctivitis*	−0.66 (−0.77 to −0.50)[c]	−0.07 (−0.28 to 0.16)	0.13 (−0.12 to 0.36)	−0.51 (−0.65 to −0.32)[c]	−0.84 (−0.88 to −0.75)[c]	1	[b]	[b]	[b]	[b]
Google USA search for *eye allergy*	0.13 (−0.06 to 0.32)	0.44 (0.24 to 0.60)[c]	0.02 (−0.22 to 0.26)	−0.43 (−0.60 to −0.22)[c]	0.05 (−0.19 to 0.28)	0.06 (−0.16 to 0.29)	1	[b]	[b]	[b]
Google USA search for *flu*	−0.15 (−0.36 to 0.05)	−0.42 (−0.59 to −0.22)[c]	−0.08 (−0.29 to 0.15)	0.65 (0.47 to 0.77)[c]	0.08 (−0.16 to 0.30)	−0.19 (−0.39 to 0.05)	−0.59 (−0.69 to −0.45)[c]	1	[b]	[b]
Google USA search for *eye drops*	0.48 (0.30 to 0.62)[c]	0.47 (0.27 to 0.62)[c]	0.04 (−0.21 to 0.27)	−0.12 (−0.34 to 0.10)	0.42 (0.20 to 0.60)[c]	−0.42 (−0.57 to −0.22)[c]	0.60 (0.42 to 0.74)[c]	−0.52 (−0.64 to −0.35)[c]	1	[b]
Twitter USA posts about *pink eye*	0.38 (0.16 to 0.56)[c]	0.09 (−0.14 to 0.31)	−0.39 (−0.56 to −0.20)[c]	0.14 (−0.12 to 0.38)	0.55 (0.33 to 0.70)[c]	−0.37 (−0.59 to −0.12)[c]	0.08 (−0.15 to 0.31)	0.05 (−0.18 to 0.27)	0.11 (−0.16 to 0.35)	1

[a] Weekly total clinical encounters at the University of California San Francisco (UCSF) for nonallergic conjunctivitis, allergic conjunctivitis, glaucoma, and influenza were compared with equivalent weekly data on keywords obtained from Google Trends and Twitter using the Spearman rank correlation.

[b] The correlation matrix is symmetric, but entries to the right are suppressed for clarity.

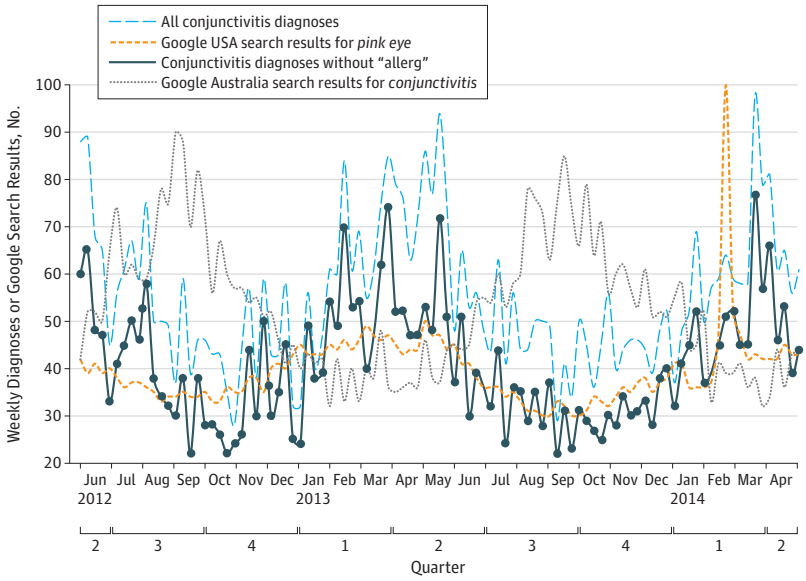
[c] $P < .001$.

results of a Google search in the United States (Google USA) for *pink eye* (ρ, 0.68 [95% CI, 0.52 to 0.78]; $P < .001$) and with Google USA search results for *eye drops* (ρ, 0.48 [95% CI, 0.30 to 0.62]; $P < .001$). We found moderate correlation between UCSF diagnoses of nonallergic conjunctivitis and Twitter USA posts about *pink eye* (ρ, 0.38 [95% CI, 0.16 to 0.56]; $P < .001$) and between UCSF diagnoses of nonallergic conjunctivitis and UCSF diagnoses of influenza (ρ, 0.33 [95% CI, 0.12 to 0.49]; $P < .001$). However, we found no strong evidence of correlation between UCSF diagnoses of nonallergic conjunctivitis and UCSF diagnoses of allergic conjunctivitis (ρ, 0.21 [95% CI, −0.02 to 0.42]; $P = .06$), between UCSF diagnoses of nonallergic conjunctivitis and Google USA search results for *eye allergy* (ρ, 0.13 [95% CI, −0.06 to 0.32]; $P = .19$), or between UCSF diagnoses of nonallergic conjunctivitis and UCSF diagnoses of glaucoma. Finally, we found evidence of inverse correlation of UCSF diagnoses of nonallergic conjunctivitis and Google Australia search results for *conjunctivitis* (ρ, -0.66 [95% CI, −0.77 to −0.50]; $P < .001$), suggesting somewhat opposite seasons, as is known for the 2 hemispheres. Similar to UCSF diagnoses of nonallergic conjunctivitis, Google USA search results for *pink eye* also correlated strongly with Google USA search results for *eye drops* (ρ, 0.42 [95% CI, 0.20 to 0.60]; $P < .001$) and with Twitter USA posts about *pink eye* (ρ, 0.55 [95% CI, 0.33 to 0.70]; $P < .001$), had a strong inverse correlation with Google Australia search results for *conjunctivitis* (ρ, -0.84 [95% CI, −0.88 to −0.75]; $P < .001$), and had no correlation with Google USA search results for *eye allergy*.

For patients at UCSF with allergic conjunctivitis, there was evidence of a strong correlation with Google USA search results for *eye drops* (ρ, 0.47 [95% CI, 0.27 to 0.62]; $P < .001$) and with Google USA search results for *eye allergy* (ρ, 0.44 [95% CI, 0.24 to 0.60]; $P < .001$) (Table 2). However, UCSF diagnoses of allergic conjunctivitis were inversely correlated with Google USA search results for *flu* (ρ, -0.42 [95% CI, −0.59 to −0.22]; $P < .001$) and UCSF diagnoses of influenza (ρ, -0.30 [95% CI, −0.49 to −0.08]; $P < .001$). We found modest correlation of UCSF diagnoses of allergic conjunctivitis with UCSF diagnoses of glaucoma (ρ, 0.24 [95% CI, 0.02 to 0.44]; $P = .02$). We found no evidence of correlation of UCSF diagnoses of allergic conjunctivitis with UCSF diagnoses of nonallergic conjunctivitis and with Google USA search results for *pink eye*, Google Australia search results for *conjunctivitis*, or Twitter USA posts about *pink eye*. Similar to UCSF diagnoses of allergic conjunctivitis, we also found evidence that Google USA search results *for eye allergy* were strongly correlated with Google USA search results for *eye drops* (ρ, 0.60 [95% CI, 0.42 to 0.74]; $P < .001$), but inversely correlated with Google USA search results for *flu* (ρ, -0.59 [95% CI, −0.69 to −0.45]; $P < .001$).

University of California San Francisco diagnoses of glaucoma (Table 2), a control diagnosis, did not correlate strongly with any tested data sources but did correlate inversely with

Figure 2. Selected Weekly Results of Google Searches and Conjunctivitis Diagnoses



Google USA results for *pink eye*, Google Australia results for *conjunctivitis* (apparent inverse seasonality), diagnoses of nonallergic conjunctivitis (those without the string "allerg" in the electronic medical record), and all conjunctivitis diagnoses.

Twitter USA posts about *pink eye* (ρ, -0.39 [95% CI, −0.56 to −0.20]; $P < .001$) and Google USA search results for *pink eye* (ρ, -0.32 [95% CI, −0.51 to −0.09]; $P < .001$). In addition, UCSF diagnoses of influenza (Table 2) correlated with Google USA search results for *flu* (ρ, 0.65 [95% CI, 0.47 to 0.77]; $P < .001$) and Google USA search results for *pink eye* (ρ, 0.47 [95% CI, 0.25 to 0.66]; $P < .001$), but correlated inversely with Google USA search results for *eye allergy* (ρ, -0.43 [95% CI, −0.60 to −0.22]; $P < .001$).

Figure 2 depicts Google USA search results for *pink eye*, Google Australia search results for *conjunctivitis*, and UCSF diagnoses of nonallergic conjunctivitis and all cases of conjunctivitis, including the apparent inverse seasonality between the United States and Australia that was also suggested in Table 2. Unlike in Figure 1, the data from UCSF are presented weekly to allow comparison with the weekly search data available from Google Trends.

Discussion

The clinical data, Google search results, and Twitter posts show a common pattern. We found evidence that clinical diagnoses of conjunctivitis detected through EMRs appear seasonal and are highly correlated with results of Google searches and correlated with relevant Tweets. We found that Google searches for *pink eye* and related terms in the United States followed the seasonality seen in prior studies.[27] Previous studies have also found that allergic rhinitis (which is related to allergic conjunctivitis), assessed through Google Trends, peaked in the spring, similar to our findings for allergic conjunctivitis.[28] This finding suggests some overlap of allergic conjunctivitis and allergic rhinitis in social media data and clinical diagnoses.[5] We also found differences in allergic vs nonallergic conjunctivitis where EMR data on nonallergic conjunctivitis correlated strongly with Google USA search results for *pink eye* but not significantly with UCSF diagnoses of allergic conjunctivitis or Google USA search results for *eye allergy*. Inversely, EMR data on allergic conjunctivitis correlated strongly with Google USA search results for *eye allergy* (and with the typical annual San Francisco area high allergy season of March through May), but not with Google USA search results for *pink eye* or EMR data on nonallergic conjunctivitis data. However, EMR data on allergic and nonallergic conjunctivitis correlated well with Google USA search results for *eye drops*. We did not find evidence that EMR data on conjunctivitis or Google USA search results for *pink eye* were correlated with glaucoma, a largely unrelated ophthalmologic condition with a reported seasonality,[27] which served as one kind of control. Electronic medical record data on influenza showed the highest correlation with Google USA search results for *flu*, as might be expected based on studies of influenza-like illness.[31] Electronic medical record data on influenza were also correlated with both EMR data on nonallergic conjunctivitis data and Google USA search results for *pink eye* (probably owing to similar seasonality of the underlying infections). Electronic medical record data on influenza were inversely correlated with EMR data on allergic conjunctivitis and Google USA search result for *eye allergy*, perhaps owing to the fact that the allergy season in the San Francisco area does not coincide with the typical influenza season. For influenza, disease-related data from search engines and social media can reveal facets of the true epidemiologic factors of the disease; in this case, we have found that to be likely for conjunctivitis, including possibly by nonallergic and allergic subtype. This finding suggests that data from search engines and social media could serve as a surrogate source of epidemiologic information about infectious eye disease, at a minimum to better refine estimates of US seasonality, but we believe this work must be conducted and validated carefully, leveraging the complementary aspects of these data vs EMR data when possible. Although EMR data may be more costly and access to data more delayed, it most likely will remain the most precise source, for example, to distinguish demographics or subtype of

eye disease and perhaps may include some diseases less likely to be identifiable online owing to nonunique search terms or to stigma (eg, rare sexually transmitted infections affecting the eye). On the other hand, social media, search engine, and other similar online sources of large amounts of nontraditional data, especially when a disease with unique keywords (such as *pink eye*) can complement this precision through more publicly (and rapidly) available data, across wider geographic regions at potentially lower cost compared with EMR data and may have an advantage of reflecting disease that does not always appear in a clinical setting (such as mild conjunctivitis). With better refinement, perhaps as with influenza and other infectious diseases, our findings also could be interpreted to suggest that social media and search engines could potentially be leveraged to identify or even predict infectious eye disease, but whether this use is possible or practical remains to be explored, including key issues such as how to distinguish seasonal increases from localized outbreaks or how to improve other desired aspects of data precision.

Our study had several limitations. Electronic medical records from only a relatively short time were available for our analysis; future analysis with a larger clinical sample size and longer time can test whether the observed pattern continues. Moreover, administrative data, such as diagnosis codes, may contain inaccuracies, and more refined means of segregating encounters for analysis may be available based on types of conjunctivitis in addition to the methods we used that might be useful for more in-depth future study. For influenza studies, often regularly government-reported data on influenza-like illness (rather than diagnosis) are used as a criterion standard, and for that approach these same inherent limitations likely exist. National ophthalmology registries may be more precise alternative larger sources of data for future studies, but regular government reporting does not currently exist for primary infectious eye disease, lending value to the use of social media and search engines as alternate sources of epidemiologic information on infectious eye diseases. For Tweets, although we used a refined query to reduce unrelated signals, it is likely that not all Tweets that mention *pink eye* are necessarily indicative of currently having conjunctivitis; some may reflect past episodes, hopes to avoid the disease, or other mentions. More refined queries can be developed and more specific information regarding age, sex, disease severity, or geographic location from Tweets would help improve analyses based on these demographics, but such information is not available at this time. Raw Google search data are not available; Google Trends provides data in a normalized format, and, in general, refinement to remove likely confounders is difficult (eg, in Figure 2, a spike in Google USA results for *pink eye* in February 2014 was likely related in part to a popular television sports anchor whose highly publicized case of conjunctivitis occurred during the Olympics). It also is possible that some inherent seasonality of Google search results might exist overall; however, not all search results showed the same seasonality, suggesting that our results were not driven by inherent seasonality of Google search results. Tweets regarding having pink eye showed the same associations with EMR data on conjunctivitis as did Google search results, also lending support to the validity of social media and search engine results as reflecting the seasonality of EMR data and indicating that conclusions based on one form of social media or search engine results can support those drawn from the other. For analysis on a national level, we found geolocation not likely to be a concern (as data from searches of Google Australia showed a near inverse pattern that one would expect based on their opposite seasons than the United States), but further refinements may allow analysis of social media based on more granular, reliable locations, such as state.

In analyzing our EMR clinical data, several findings were reported regarding nonallergic vs allergic encounters. We found relatively more patients with nonallergic than allergic conjunctivitis in the youngest age groups, perhaps explained by a higher incidence of bacterial or viral conjunctivitis in young children or based on a higher rate of allergy in older patients (for patients ≥50 years, there were more diagnoses of allergic than nonallergic conjunctivitis, as has been reported[3]). We also found a tendency for more conjunctivitis cases overall among females than males and by conjunctivitis subgroup (especially for allergic conjunctivitis). We also found conjunctivitis subgroup differences based on race/ethnicity. More in-depth analysis of a larger EMR data set and subsets could help to better explain some of our findings on conjunctivitis related to social media. Future analyses investigating the correlation of other eye infections and eye diseases with internet-based social media and search engines may be useful.

Conclusions

Internet-based search engine and social media data were strongly associated with the occurrence of clinically diagnosed conjunctivitis as seen in EMRs. The information that people post and search for online, and when they post such information, can be leveraged to better understand the epidemiologic factors of conjunctivitis.

ARTICLE INFORMATION

Accepted for Publication: May 20, 2016.

Published Online: July 14, 2016. doi:10.1001/jamaophthalmol.2016.2267.

Author Contributions: Drs Deiner and Porco had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: All authors.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Deiner, Porco.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Deiner, Porco.
Obtained funding: Deiner, Lietman, Porco.
Administrative, technical, or material support: Deiner, Porco.
Study supervision: Lietman, Porco.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Funding/Support: This work was 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, an unrestricted grant from Research to Prevent Blindness, through the University of California San Francisco Information Technology Enterprise Information Analytics Department's Research Data Browser and Clinical

Data Research Consultation Services, and grant 1-U01-GM087728 from the NIH–National Institute of General Medical Sciences Models of Infectious Disease Agent Study Program (Dr Porco).

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Epling J. Bacterial conjunctivitis. *BMJ Clin Evid*. 2012;2012.

2. Meyer-Rüsenberg B, Loderstädt U, Richard G, Kaulfers PM, Gesser C. Epidemic keratoconjunctivitis: the current situation and recommendations for prevention and treatment. *Dtsch Arztebl Int*. 2011;108(27):475-480.

3. Singh K, Axelrod S, Bielory L. The epidemiology of ocular and nasal allergy in the United States, 1988-1994. *J Allergy Clin Immunol*. 2010;126:778-783.e776.

4. Saban DR, Calder V, Kuo CH, et al. New twists to an old story: novel concepts in the pathogenesis of allergic eye disease. *Curr Eye Res*. 2013;38(3):317-330.

5. Leonardi A, Castegnaro A, Valerio AL, Lazzarini D. Epidemiology of allergic conjunctivitis: clinical appearance and treatment patterns in a population-based study. *Curr Opin Allergy Clin Immunol*. 2015;15(5):482-488.

6. Azari AA, Barney NP. Conjunctivitis: a systematic review of diagnosis and treatment [published correction appears in *JAMA*. 2014;311(1):95]. *JAMA*. 2013;310(16):1721-1729.

7. Kuo IC, Espinosa C, Forman M, Valsamakis A. A polymerase chain reaction-based algorithm to detect and prevent transmission of adenoviral conjunctivitis in hospital employees. *Am J Ophthalmol*. 2016;163:38-44.

8. Smith AF, Waycaster C. Estimate of the direct and indirect annual cost of bacterial conjunctivitis in the United States. *BMC Ophthalmol*. 2009;9:13.

9. Schneider JE, Scheibling CM, Segall D, Sambursky R, Ohsfeldt RL, Lovejoy L. Epidemiology and economic burden of conjunctivitis: a managed care perspective. *J Managed Care Med*. 2014;17:78-83.

10. Brownstein JS, Freifeld CC. HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveill*. 2007;12(11):E071129.5.

11. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the Web for public health surveillance. *N Engl J Med*. 2009;360(21):2153-2155, 2157.

12. Madan A, Cebrian M, Lazer D, Pentland A. Social sensing for epidemiological behavior change. *Proceedings of the 12th Association for Computing Machinery International Conference on Ubiquitous Computing*. Copenhagen, Denmark: Association for Computing Machinery; 2010:291-300.

13. Khan K, McNabb SJ, Memish ZA, et al. Infectious disease surveillance and modelling across geographic frontiers and scientific specialties. *Lancet Infect Dis*. 2012;12(3):222-230.

14. Sadilek A, Kautz H, Bigham JP. Modeling the interplay of people's location, interactions, and social ties. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*: AAAI Press; 2013:3067-3071.

15. Hartley DM, Nelson NP, Arthur RR, et al. An overview of internet biosurveillance. *Clinical Microbiol Infect*. 2013;19(11):1006-1013.

16. 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(1):7-33.

17. Nuti SV, Wayda B, Ranasinghe I, et al. The use of Google Trends in health care research: a systematic review. *PLoS One*. 2014;9(10):e109583.

18. Brownstein JS, Mandl KD. Reengineering real time outbreak detection systems for influenza epidemic monitoring. *AMIA Annual Symposium Proceedings* 2006;866.

19. Brownstein JS, Freifeld CC, Madoff LC. Influenza A (H1N1) virus, 2009—online monitoring. *N Engl J Med*. 2009;360(21):2156.

20. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012-1014.

21. 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(3):e90536.

22. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. *PLoS Comput Biol*. 2014;10(11):e1003892.

23. 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(10):e1004513.

24. Hoen AG, Keller M, Verma AD, Buckeridge DL, Brownstein JS. Electronic event-based surveillance for monitoring dengue, Latin America. *Emerg Infect Dis*. 2012;18(7):1147-1150.

25. McGregor F, Somner JE, 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(1):46-52. doi:10.1111/opo.12093.

26. HealthMap. HealthMap. http://www.healthmap.org/. Accessed December 12, 2014.

27. Leffler CT, Davenport B, Chan D. Frequency and seasonal variation of ophthalmology-related internet searches. *Can J Ophthalmol*. 2010;45:274-279.

28. Kang MG, Song WJ, Choi S, et al. Google unveils a glimpse of allergic rhinitis in the real world. *Allergy*. 2015;70(1):124-128.

29. Crimson Hexagon. Social data you can work with. http://www.crimsonhexagon.com/. Accessed June 13, 2016.

30. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall; 1993.

31. Dugas AF, Hsieh YH, Levin SR, et al. Google flu trends: correlation with emergency department influenza rates and crowding metrics. *Clin Infect Dis*. 2012;54(4):463-469.

Invited Commentary

The Utility of "Big Data" and Social Media for Anticipating, Preventing, and Treating Disease

Alfred Sommer, MD, MHS

Our new era of clinical practice is largely the consequence of profound advances in imaging, devices, and therapeutics. But 2 other developments, outside our traditional domain, are beginning to affect our ability to predict, prepare for, and treat disease: "big data" and social media. The influence that big data and social media have had has been mostly tentative to date, as we learn how best to deploy them. Big data encompasses the increasing ubiquity of electronic medical records (EMRs) and other data sources that are potentially linked via the internet, such as sales of over-the-counter medications and records of school absenteeism, that enable real-time syndromic surveillance of instances of disease. A wide variety of social media, from Google searches to Tweets, may reflect the occurrence of such instances of disease.

Related article page 1024

For many diseases, EMR data and other syndromic data are too often restricted to a single institution or small number of institutions with compatible systems. The utility of data from social media is limited by ethnic-, sex-, and age-specific pref-