AMERICAN ACADEMY OF OPHTHALMOLOGY®

Google Searches and Detection of Conjunctivitis Epidemics Worldwide

Michael S. Deiner, PhD,[1,2] Stephen D. McLeod, MD,[1,2] Jessica Wong, MS,[2] James Chodosh, MD, MPH,[3] Thomas M. Lietman, MD,[1,2,4] Travis C. Porco, PhD, MPH[1,2,4]

Purpose: Epidemic and seasonal infectious conjunctivitis outbreaks can impact education, workforce, and economy adversely. Yet conjunctivitis typically is not a reportable disease, potentially delaying mitigating intervention. Our study objective was to determine if conjunctivitis epidemics could be identified using Google Trends search data.

Design: Search data for conjunctivitis-related and control search terms from 5 years and countries worldwide were obtained. Country and term were masked. Temporal scan statistics were applied to identify candidate epidemics. Candidates then were assessed for geotemporal concordance with an a priori defined collection of known reported conjunctivitis outbreaks, as a measure of sensitivity.

Participants: Populations by country that searched Google's search engine using our study terms.

Main Outcome Measures: Percent of known conjunctivitis outbreaks also found in the same country and period by our candidate epidemics, identified from conjunctivitis-related searches.

Results: We identified 135 candidate conjunctivitis epidemic periods from 77 countries. Compared with our a priori defined collection of known reported outbreaks, candidate conjunctivitis epidemics identified 18 of 26 (69% sensitivity) of the reported country-wide or island nationwide outbreaks, or both; 9 of 20 (45% sensitivity) of the reported region or district-wide outbreaks, or both; but far fewer nosocomial and reported smaller outbreaks. Similar overall and individual sensitivity, as well as specificity, were found on a country-level basis. We also found that 83% of our candidate epidemics had start dates before (of those, 20% were more than 12 weeks before) their concurrent reported outbreak's report issuance date. Permutation tests provided evidence that on average, conjunctivitis candidate epidemics occurred geotemporally closer to outbreak reports than chance alone suggests ($P < 0.001$) unlike control term candidates ($P = 0.40$).

Conclusions: Conjunctivitis outbreaks can be detected using temporal scan analysis of Google search data alone, with more than 80% detected before an outbreak report's issuance date, some as early as the reported outbreak's start date. Future approaches using data from smaller regions, social media, and more search terms may improve sensitivity further and cross-validate detected candidates, allowing identification of candidate conjunctivitis epidemics from Internet search data potentially to complementarily benefit traditional reporting and detection systems to improve epidemic awareness. *Ophthalmology 2019;126:1219-1229 © 2019 by the American Academy of Ophthalmology*

Supplemental material available at www.aaojournal.org.

Big data and web-based surveillance have been applied to infectious disease surveillance.[1–15] It has been suggested that such efforts may complement traditional reporting or provide insight into infectious conditions that may be underreported or generally not reportable. These include conjunctivitis,[16] a condition only reportable in the United States for neonatal cases despite substantial economic[17] and public health[18,19] impact and for which it has been shown that early public awareness has potential to improve outcomes.[20]

Evidence has been reported previously that available online search and social media data sources—including tweets, blog posts, forums, and search engine query data—reflect the age- and cause-specific features of the clinical epidemiologic aspects and seasonal patterns of conjunctivitis.[21–24] In this study, we tested the hypothesis that Internet search data for key words relevant to conjunctivitis could be used to identify actual conjunctivitis epidemics. Specifically, we tested whether, while masked, we could identify candidate epidemics of conjunctivitis. We validated these identifications using reports of conjunctivitis outbreaks after unmasking. We also assessed the outcomes overall, as well as for countries individually, and when countries are grouped within their Global Burden of Diseases (GBD) regions based on closeness geographically and epidemiologically.[25]

Methods

In this section, we describe (1) how we obtained Google search data for identifying epidemics, (2) how we identified apparent ("candidate") epidemics from these time series, (3) how we identified actual reports of known conjunctivitis outbreaks, and (4) how

https://doi.org/10.1016/j.ophtha.2019.04.008
ISSN 0161-6420/19

we validated our detection method using those reports. All research adhered to the tenets of the Declaration of Helsinki. The Institutional Review Boards approved the study and waived the requirement for informed consent because of the retrospective nature of the study.

Google Search Data for Identifying Epidemics

The Google Timeline for Health application programming interface allows researchers and others, after applying to Google and being granted permission, to access Google Trends data regarding the geotemporal location of online searches. These data have been used, for example, to study behavior, to explore outbreaks, and to forecast economic activity.[26,27] Using this application programming interface, we collected worldwide, national-level, daily Google search data for 24 key words. These key words included terms related to conjunctivitis in several languages, as well as positive control key words related to other diseases and negative control key words designed to reflect general changes in search data volume (to account for any search volume changes presumably unrelated to disease). For a list of key words, please see Supplement VI (available at www.aaojournal.org). Data were obtained from July 10, 2012, through July 9, 2017. The resulting time series represents relative search interest data, reflecting the proportion of searches for the term of interest from among all searches for all terms for a given geography and period. The proportion is calculated by Google using a random sample. Very small values are censored by the application programming interface, partly to protect privacy; such censored values appear as 0s in the time series. Country-search term combinations yielding no relative search information were excluded.

Identification of Candidate Epidemics

To identify epidemics retrospectively, we used an approach that implemented 3 variants of the scan statistic[28] in an automated fashion to all time series. The first algorithm ("Scan") was applied to all data and used a modified temporal scan statistic based on first applying a 5-day centered median filter after linear detrending to remove short spikes potentially resulting from media coverage. It then examined a 31-day centered moving average. The second algorithm ("Lush") then was implemented automatically but only when at least 75% of the data were available (nonzero). This regression procedure used negative binomial regression with cyclic basis splines to represent an arbitrary seasonal background distribution.[29,30] It also included quadratic secular terms. Temporal scanning then was applied to the residuals, identifying intervals when the observed value consistently exceeded the model prediction. Finally, for situations where the second method was not implemented, a third algorithm ("Sparse") was applied that first dichotomized each value in the time series, with 1 denoting at nonmissing value. It then applied a 31-day centered moving average to this series of binary values. For all methods, permutation then was used to determine the quantiles of the distribution of the expected maximum value of the moving average under the null hypothesis of a stationary series. Epidemic detection thus was accomplished by examining candidates from this automatically applied collection of algorithms. Further details are provided in Supplement V (available at www.aaojournal.org). For each epidemic, the first date at which the threshold was exceeded was considered the earliest detectable date, that is, the start date.

Statistical identification of candidate epidemics was conducted in a masked manner by concealing the search term and location (country) from the data analyst. Larger-scale geographic information also was not used. After all candidate epidemics were identified, masked terms and geolocation were unmasked for subsequent validation comparisons with reported outbreaks.

Conjunctivitis Outbreak Reports: Program for Monitoring Emerging Diseases, PubMed, and Other Online Sources

Our study was designed to identify conjunctivitis outbreaks that, in many countries, are not reported in any standardized or systematic manner. Although a true gold standard therefore is not available, other sources of reports can validate our candidate epidemics. In late summer 2017, we identified conjunctivitis outbreak reports from July 10, 2012, through August 9, 2017, using 3 sources: Program for Monitoring Emerging Diseases (ProMED), PubMed, and other online Internet content. The ProMED, sponsored by the International Society for Infectious Diseases, is an Internet-based system allowing rapid reporting and dissemination of information on infectious disease outbreaks worldwide, including conjunctivitis. The ProMED has been an early warning system for infectious diseases for more than 22 years.[31–34] We queried the ProMED and PubMed for conjunctivitis outbreak reports using their online search tools. We used a standardized query to locate additional online reports (news stories and other Internet content) of human conjunctivitis outbreaks. Supplement VI includes details of queries we used to identify outbreak reports from these 3 sources.

For all reports of conjunctivitis outbreaks, we recorded the report issuance date, reported start date, and country. We categorized each outbreak as country-wide, island nation-wide, or both; region-wide, district-wide, or both; nosocomial; or small (e.g., 1 classroom, but not associated with a health care facility). We excluded reports with unclear start dates (less precise than a 1-month window) or start dates not occurring between July 18, 2012, and July 2, 2017. We excluded 1 report later identified as a hoax. These data were not revealed to team members conducting masked candidate detection until after they completed candidate identification analysis.

A given outbreak in a specific country may be documented in multiple reports. Similarly, for a given outbreak, multiple candidate epidemics identified from analysis of Google search data may be close together in time. To compare 3 report sources and identify candidates for each outbreak, we identified a single start date for each of these data sources per outbreak. More specifically, for each country, we recorded the earliest reported outbreak start date. The period from the first start date to 45 days after the last report date was considered a window of interest that we refer to as a 45-day continuum period. If 46 days or more separated consecutive reports in the same country, we considered a new epidemic (new continuum identification period) to have begun with the second report. This resulted in a defined set of 45-day continuum periods for each country, which we used for sensitivity and validation analysis.

Validation of Candidate Epidemic Detection

Overall, for all countries combined, we estimated the sensitivity of our candidate epidemic detection in 2 ways. First, for each reported outbreak in a country, we determined whether at least 1 candidate epidemic was identified within the same continuum. Second, we repeated the analysis for the 4 categorized outbreak sizes. For comparison, we also assessed sensitivity by tabulating the frequency of windows of interest containing overlap with a statistical detection window. In this second approach, we considered an epidemic to be occurring for 31 days after each identified start date, based on all candidate detection algorithms. For comparison, we compared candidate epidemics identified from control terms with

conjunctivitis report dates. Confidence intervals for proportions were computed using the exact (Clopper-Pearson) method.

In the absence of a gold standard, the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of our candidate epidemics, strictly speaking, cannot be computed. However, on a country-based level, we estimated a measure of these quantities based on the following assumptions: (1) the 40 total possible 45-day continuum periods (1825 days/45 days) in the time series for a country serve as an effective denominator of 40 (a maximum possible of 40 continuum periods per country that each could have had a candidate in them); (2) any continuum period containing a start date for a reported outbreak was considered to reflect a true epidemic, and any not containing one was assumed to have experienced no epidemic; and (3) any continuum period containing a start date for a candidate conjunctivitis epidemic effectively tested positive, whereas any continuum period not containing a candidate corresponded to a negative test result. All results were calculated first for countries containing reported outbreaks. Countries in which there was no reported outbreak at all, but for which we found candidate epidemics, then were grouped into similar regions, defined using the Institute for Health Metrics and Evaluation (IHME) GBD 2016 location hierarchy file as a guide[25,35] to assess whether false-positive results (candidates identified in countries with no outbreak reports) may have been validated from an outbreak reported for a nearby similar country. Countries with no candidates were assessed only for specificity and NPV. The mean sensitivity and other summary values of all the country-level results also were computed as a secondary comparison with our other all-country analyses.

For all reported outbreaks for which we identified candidate conjunctivitis epidemics within the same 45-day continuum and country, we assessed the number of days between the start date described in the outbreak report and the start date of the identified candidate epidemic. Overall, we also assessed the portion of detected candidates with an earlier versus later start date than the start dates described in their corresponding outbreak reports. In a similar manner, we also compared the issuance dates (first appearance in print) of the outbreak reports with the start dates of the corresponding identified candidate conjunctivitis epidemics.

For all countries combined, we also assessed the statistical association of candidate epidemics with reported outbreaks based on a simple permutation test. We permuted a country at random, and within countries, we conducted a random cyclic permutation of the starting times (accounting for the nonindependent nature of the time series of both reported outbreaks and detected candidates). The test statistic was the number of candidate epidemic starting dates that fell within continuum period regions of interest. If more of the candidate start dates fell within windows of interest than expected by chance alone, we rejected the null hypothesis of no association when $P < 0.05$. As a control, using candidates we had identified (while masked) from the negative control non-conjunctivitis search terms, we repeated the permutation test assessment.

In addition, we computed the frequency of nonepidemic days (days falling outside our continuum periods) that did not overlap a detection window as an alternate measure of specificity. We also assessed whether detection times of candidates identified for negative control terms (*for*, *para*) differed statistically with detection times based on conjunctivitis terms using PERMANOVA. Finally, we classified each day in our 5-year period as being part of a candidate epidemic or not. Using these binary series, we compared the φ correlation based on conjunctivitis search terms and search terms related to influenza, allergy, and negative control (*for* and *para*) terms.

Institutional Review Board Approval

The University of California, San Francisco, Institutional Review Board approved the study before it commenced (approval no.: 14-14743).

Results

Outbreak Reports

A priori, before any comparison with candidate epidemics, we identified 87 conjunctivitis outbreak reports from 49 countries from July 17, 2012, through July 2, 2017. We excluded reports from countries yielding no Google search information (3 ProMED, 1 PubMed, and 4 other online outbreak reports; see "Discussion"). Within each 45-day continuum, we then selected only the first occurrence of each report source (dropping duplicate reports of the same outbreak if from the same source), resulting in 20 ProMED reports from 18 countries, 7 PubMed reports from 7 countries, and 37 other online reports from 27 countries. If 1 continuum contained multiple reports, for sensitivity analysis only, one of these was used for that continuum. All reports were used when deriving date difference comparisons. This final set of outbreak reports was used for comparison with our candidate epidemics (Table 1).

Candidate Outbreaks Detected from Conjunctivitis-Related Search Terms and Scan Methods

From all search terms combined, we identified 1166 candidate epidemics, of which 293 were from search terms representing conjunctivitis. These conjunctivitis candidates from different search terms, scan methods, or both often were close in time. We selected the first from each 45-day continuum period, resulting in a final set of 135 candidate epidemic continuum periods from 77 countries.

The 3 most common first conjunctivitis search terms within continuums were *conjunctivitis* (n = 43 [32% of total]), *conjonctivite* (n = 26 [19% of total]), and *conjuntivitis* (n = 26 [19% of total]). For some conjunctivitis search terms specific to certain locations, we often detected only epidemics in those locations. For example, conjunctivitis candidates detected from the conjunctivitis search term *aankh aana* (Hindi) were found 6 times and only in India, those for term *apollo eye* were found 4 times and only in Nigeria, and those for term *azoumounou* (Haitian creole) were found only for Haiti (3 times) and the United States (once). More details of the results for all search terms can be seen in Supplement I, Tables S3, S4, and S5 (available at www.aaojournal.org).

Analyzing the success of the 3 scan methods used, for all search terms combined, we found 76% of candidates were identified using Scan, 18% using Sparse, and 5% using Lush. For a more detailed comparative analysis and visualization of the results from the 3 scan methods used, please see Supplement II and Figure S3 (available at www.aaojournal.org).

Detected Candidate Epidemic Concurrence with Reported Outbreaks

In Tables 1 and 2, rows in black indicate reported outbreaks that validated candidate epidemics within the same 45-day outbreak continuum period. These tables also allow a comparison of day differences between candidate conjunctivitis epidemic start dates (the leading edge of the scan window for which the epidemic threshold was reached initially) and the start date of each reported outbreak in the same 45-day continuum and allows a comparison

Table 1. Country-Wide and Island Nation-Wide Outbreak Reports Compared with Identified Candidate Epidemics

Country*	Found†	Report Source‡	Reported Start§	Days before Start‖	Report Issuance¶	Days before Issuance#	Report Reference**
American Samoa	Yes	Other	2014-04-01	−6	2014-04-09	2	1a
Antigua & Barbuda	Yes	Other	2017-06-15	−24	2017-07-04	−5	1b
Bahamas	Yes	ProMED	2017-05-15	−15	2017-06-20	21	1c
Burkina Faso	Yes	ProMED	2016-08-15	11	2016-09-07	34	1d
Cambodia	Yes	Other	2013-10-04	3	2013-10-25	24	1e
Cuba	No	ProMED	2017-07-01		2017-07-29		1f
Dominican Republic	Yes	ProMED	2017-05-06	−7	2017-05-27	14	1g
Fiji	Yes	ProMED	2016-03-15	−9	2016-04-01	8	1h
France	No	ProMED	2017-05-20		2017-06-24		1i
Guadeloupe	Yes	ProMED	2017-05-14	−3	2017-06-08	22	1j
Guam	Yes	ProMED	2014-05-15	−22	2014-06-03	−3	1k
Haiti	Yes	Other	2017-05-15	35	2017-05-15	35	1l
Honduras	Yes	Other	2017-06-07	10	2017-07-25	58	1m
Martinique	Yes	ProMED	2017-05-14	−20	2017-06-08	5	1n
Mauritius	Yes	Other	2015-02-23	20	2015-03-15	40	1o
Mauritius	No	Other	2016-04-11		2016-05-03		1p
Nicaragua	No	ProMED	2013-01-01		2013-02-21		1q
Réunion	Yes	PubMed	2015-01-15	−41	2016-06-26	487	1r
Samoa	Yes	Other	2014-03-15	−3	2014-03-25	7	1s
Singapore	No	Other	2014-09-07		2014-09-07		1t
Somalia	No	ProMED	2014-12-01		2014-12-07		1u
Thailand	No	Other	2014-01-01		2014-02-21		1v
Thailand	Yes	PubMed	2014-07-01	−41	2015-03-31	232	1w
Thailand	No	Other	2016-05-01		2016-06-05		1x
Tonga	Yes	Other	2016-05-01	−12	2016-10-11	151	1y
Vietnam	Yes	Other	2013-09-01	−15	2013-09-20	4	1z

ProMED = Program for Monitoring Emerging Diseases.
Rows corresponding to the reported country-wide and island nation-wide outbreaks that were detected with Google search data are shown in black; others are shown in gray. The first report from each possible report source within each 45-day continuum period is shown and compared with dates and locations of identified candidate epidemics (note, only 1 report per continuum, that with the earliest reported start date was used in calculating sensitivity in this study).
*Name of the country.
†Whether candidate was found, in a masked fashion, within the same 45-day continuum as the reported outbreak.
‡The source from where outbreak reports were obtained using queries of ProMED, PubMed, and other online (other) reports.
§The start date of the outbreak defined in the report.
‖The candidate start date's number of days before the report's reported start date, if within same continuum (a positive number of days indicates the candidate start date occurred that many days before the report's reported start date).
¶The date the report was published.
#The candidate start date's number of days before the report's issuance date, if start dates were within same continuum (a positive number of days indicates the candidate start date was that many days before the report's issuance date).
**The cited original source of the reported outbreak. See Supplement 1 for Table 1 Outbreak Report references.

with the report's issuance date. More detailed analysis of these results, including sensitivity, start or report date differences, frequently validated key words, and percentage of candidates validated by reports is described next (as well as in the Supplemental materials I-IV, VI). Daily searches, detected candidates, and outbreak reports are compared visually in time series in Figures 1 and 2. Figure 1 shows examples for 5 countries of time series search data (rows 2 onward) for a number of conjunctivitis-related and control terms, and any candidate epidemics identified are shown as red triangles with their identified start date, with the top row showing outbreak reports as inverted gold triangles. Figure 2 shows resulting candidates detected and outbreak reports for all countries in which an outbreak report was found. Sequential triangle border colors indicate unique 45-day continuum periods, including when reports and candidates occurred within the same continuum (same border color) for a country. For some reported outbreaks, if the issuance date of the report occurred 1 week or more after the plotted reported start date, a dotted grey line leads to the right of the gold triangle to indicate the issuance date.

Worldwide Validations by Outbreak Size Using the Program for Monitoring Emerging Diseases, PubMed, and Other Online Outbreak Reports

Our method identified 28 of 56 (50% sensitivity; 95% confidence interval [CI], 36%−63%) of the reported outbreaks (Fig 2; Tables 1 and 2). We identified 18 of 26 (69% sensitivity; 95% CI, 48%−86%) reported country-wide or island nation-wide outbreaks, or both; 9 of 20 (45% sensitivity) reported region-wide or district-wide outbreaks, or both; 1 of 4 (25% sensitivity) reported nosocomial outbreaks; and 0 of 6 (0% sensitivity) reported small outbreaks. Although we chose to use a 45-day continuum period for our main analyses, we conducted several alternate approaches for comparison. First, we repeated the analysis above, but based on 31- or 60-day continuum periods, and found no sensitivity differences from those reported above for when using a 45-day period. Second, using an alternative time-based approach to assess sensitivity, we found similar sensitivity results as with the approaches described above (50% overall [95% CI, 37%−63%];

Table 2. Smaller Reported Outbreaks Compared with Identified Candidate Epidemics

Country*	Found†	Report Source‡	Reported Start§	Days before Start‖	Report Issuance¶	Days before Issuance#	Report Reference**
District-wide or region-wide outbreak size group							
Brazil	No	ProMED	2017-05-18		2017-06-21		2a
Costa Rica	Yes	Other	2017-06-30	3	2017-06-30	3	2b
Dominica	Yes	Other	2017-05-31	−32	2017-05-31	−32	2c
Ghana	Yes	Other	2016-07-18	34	2016-08-10	57	2d
Guyana	Yes	Other	2017-06-23	24	2017-07-15	46	2e
India	No	Other	2012-08-09		2012-08-09		2f
India	No	Other	2013-07-25		2013-08-18		2g
India	No	Other	2013-11-15		2014-05-07		2h
India	Yes	Other	2014-09-04	−16	2014-09-04	−16	2i
India	Yes	Other	2017-03-27	−18	2017-03-27	−18	2j
Mexico	No	ProMED	2017-04-09		2017-04-13		2k
Nigeria	Yes	Other	2016-10-02	−16	2016-10-02	−16	2l
Oman	No	ProMED	2014-02-15		2014-03-13		2m
Philippines	Yes	Other	2015-08-27	13	2015-08-27	13	2n
Sri Lanka	Yes	Other	2015-06-01	1	2015-06-08	8	2o
United States	No	ProMED	2012-08-09		2012-08-24		2p
Viet Nam	No	ProMED	2012-08-06		2012-08-09		2q
Viet Nam	No	Other	2014-09-15		2014-11-10		2r
Viet Nam	No	Other	2016-06-20		2016-07-05		2s
Viet Nam	No	ProMED	2017-02-10		2017-02-14		2t
Nosocomial outbreak size group							
Singapore	No	Other	2015-10-15		2015-12-11		2u
Turkey	No	PubMed	2015-01-01		2016-09-01		2v
United Kingdom	Yes	PubMed	2015-02-06	−12	2016-05-01	438	2w
United States	No	PubMed	2015-08-15		2016-04-01		2x
Small outbreak size group							
China	No	PubMed	2012-10-05		2014-10-24		2y
Hungary	No	Other	2013-10-07		2013-10-07		2z
Italy	No	ProMED	2013-08-25		2013-09-02		2aa
Sudan	No	Other	2016-03-11		2016-03-18		2bb
Uganda	No	Other	2017-01-05		2017-01-05		2cc
United States	No	Other	2016-07-20		2016-07-20		2dd

ProMED = Program for Monitoring Emerging Diseases.
All remaining outbreak reports not shown in Table 1 (i.e., those with an a priori assigned size category smaller than country-wide and island nation-wide outbreak). Rows corresponding to outbreaks that were detected with Google search data are shown in black and others are shown in gray, and the first report from each possible report source within each 45-day continuum period is shown. See Supplement 1 for Table 1 Outbreak Report references.
*Name of the country.
†Whether candidate was found, in a masked fashion, within the same 45-day continuum as the reported outbreak.
‡The source from where outbreak reports were obtained using queries of ProMED, PubMed, and other online (other) reports.
§The start date of the outbreak defined in the report.
‖The candidate start date's number of days before the report's reported start date, if within same continuum (a positive number of days indicates the candidate start date occurred that many days before the report's reported start date).
¶The date the report was published.
#The candidate start date's number of days before the report's issuance date, if start dates were within same continuum (a positive number of days indicates the candidate start date was that many days before the report's issuance date).
**The cited original source of the reported outbreak.

by size: 68% country-wide, island nation-wide, or both; 45% region-wide, district-wide, or both; 9% small and nosocomial). As a control, in contrast to results above for conjunctivitis candidates, for epidemic candidates identified (while masked to terms) from negative control terms, we found much lower overall sensitivity (7.0%; 95% CI, 1.9%−17%) when comparing with the reported outbreaks.

Validations by Country

We also analyzed results on a country level. See Supplement III, Table S6 (available at www.aaojournal.org) for individual country-level results, including specificity, NPV, false-positive count, and (for 42 countries with reported outbreaks) sensitivity and PPV. Overall, the mean specificity per country from all 149 countries combined was 0.98 (median, 1; standard deviation [SD], 0.03; minimum, 0.81; maximum, 1), mean NPV per country was 0.995 (median, 1; SD, 0.014; minimum, 0.91; maximum, 1), and the mean number of false-positive continuums per country was 0.67 (median, 0; SD, 1.18; minimum, 0; maximum, 7). For just the 42 countries with any reported outbreaks, specificity, NPV, and mean number of false-positive continuums were similar (means of 0.98, 0.98, and 0.74, respectively), and for those 42 countries, the overall mean sensitivity was 0.58 (median, 1; SD, 0.48; minimum, 0; maximum, 1), and the mean adjusted (assigning 0 if no candidates were found) PPV was 0.55. For countries with no reported outbreaks, the mean number of false-positive continuums per country was 0.64. When grouping countries by GBD region, we

Figure 1. Graph of 5 illustrative countries demonstrating daily search data, candidate epidemics identified from that data, and reported outbreaks. For each country (column), the time span provided is from the earliest to latest occurring candidate conjunctivitis epidemic or reported epidemic within the full study period (i.e., the first 4 countries shown had only a single continuum period containing any candidates or reports). The center of each point represents the start date for candidate epidemics and the issuance date for reports. Search terms not shown if all 5 countries had no available daily relative search interest values. Y-axis for each time series is normalized (% of maximum value) daily search values. Daily values are indicated by colored vertical bars; Program for Monitoring Emerging Diseases, PubMed, and other online reports by large gold inverted triangles. Conjunctivitis candidates identified from conjunctivitis-related search terms are shown by red triangles.

Figure 2. Graphical representation of time series of reported outbreaks compared with detected conjunctivitis candidate epidemic dates. For each country, conjunctivitis candidate epidemics (red-filled triangles) are plotted based on their start dates, and any reported outbreaks (gold-filled inverted triangles) for that country are plotted based on the reported start date of the report. The center of each point represents the actual dates. Each new continuum period within a country corresponds to a different triangle border color for the outbreak reports and candidate epidemics; triangles with identical border color represent reports, candidates, or both occurring within the same continuum period. For all reported outbreaks that had an issuance (publication or first online) date that was 1 week or more after that report's reported start date, a dotted black line leads to a vertical black line indicating the report's issuance date. Note: some reported start dates (used to identify continuum identification periods and to compare with candidates) were much earlier than when the report actually was issued (e.g., see Réunion, Tonga). Countries with no reported outbreaks are not shown. Gold inverted triangles represent issuance date of Program for Monitoring Emerging Diseases, PubMed, and other online reports; red triangles represent candidate conjunctivitis epidemics identified in this study from Google search term data. Border colors represent unique 45-day continuums in a country's time series. Minor breaks: 1 month.

also found that for many countries with no gold standard with which to compare, there was a reported outbreak in a neighboring country within their continuum period within that GBD region (see examples in Supplement IV, Fig S4, available at www.aaojournal.org). Therefore, for our analysis and in the table, we adjusted the sensitivity results for such countries. With this adjustment (considering neighboring country-reported outbreaks within a GBD region as a gold standard-confirmed positive test result), the overall false-positive rate improved to a mean of 0.49 continuums (of a maximum of 40 possible continuums) per country.

Start Date Comparisons

We compared our candidate epidemic start dates with the reported start dates of outbreak reports. For all sizes of epidemics combined, of the 35 reported outbreaks that our candidate detection methods identified in the same country, 13 had a reported start date that was later than the start date identified for our concurrent matching candidate epidemic (37% of our concurrent candidates had start dates before their matching reported outbreak start date). A total of 11% of the candidate start dates were 1 to 3 weeks before their matching outbreak report's reported start date, and 11% were 4 to 6 weeks before their matching outbreak report's reported start date.

Report Issuance Date Comparisons

When comparing our candidate epidemic start dates with the report issuance dates of the matching 35 outbreak reports that our candidate detection methods identified in the same country, 29 reports were issued after the start date identified for our concurrent matching candidate epidemic (83% of our concurrent candidates had start dates before their matching reported epidemic's report issuance date). Of those, 20% of our candidate start dates were 1 to 3 weeks before their matching report's issuance date, 17% were 4 to 6 weeks before their matching report's issuance date, 9% were 7 to 12 weeks before their matching report's issuance date, and 20% were more than 12 weeks before their matching report's issuance date.

Additional Validations, Including for Candidates Not Identified from Reports

When we compared the association of the candidate epidemics identified from conjunctivitis search terms (as well as those identified from control terms) with the reported observed outbreak times, using a simple permutation test, we found evidence that the candidate epidemics are closer, on average, to reported outbreaks than chance alone would suggest ($P < 0.001$, permutation test). We found no evidence that negative controls yield candidate epidemics that are closer to reported conjunctivitis outbreaks than chance alone ($P = 0.40$). A measure of specificity also was computed by determining the fraction of nonepidemic days that are not 31 days after an identified candidate epidemic. For conjunctivitis candidates, this value was 95.4% (95% CI, 94.0%−96.2%). A similar result was found for negative control term candidates.

Comparing candidates identified for conjunctivitis search terms with those identified from negative control search terms, we found evidence that the timing of these detected epidemics was different using a permutation PERMANOVA test ($P < 0.001$). Similarly, φ correlations found little evidence for correlation between these 2 search term candidate epidemic groups (φ = 0.04; 95% CI, 0.01−0.08) or between conjunctivitis term candidates and allergy term-related candidates (φ = 0.09; 95% CI, 0.03−0.18) or influenza (φ = 0.05; 95% CI, 0.01−0.11).

Discussion

Overall, our study found evidence that scan statistics conducted on Google search data yield informative candidate epidemics. Continuous monitoring for conjunctivitis outbreaks in many countries around the world in near real time using search data may be possible, complementing identification of conjunctivitis outbreaks detected from clinical monitoring systems. Studies have shown value in the use of multiple data sources to identify and take action better in response to outbreaks, including for conjunctivitis.[19] Although the issuance date of some formal public health agency reports presumably can lag simply because of administrative delays, in some cases, delays may be the result of limited resources for identifying or confirming suspected outbreaks. It may become possible to notify such agencies early about a likely conjunctivitis candidate epidemic before the date that a public health report would be issued and potentially to accelerate awareness, confirmation, and official public health reporting. For infectious epidemics, studies suggest that there is a benefit of reducing the impact of outbreaks,[36] including through social distancing reducing transmission, such as for influenza.[37] Some evidence suggests that early public warning improves conjunctivitis outcomes and that conjunctivitis surveillance and public awareness may improve clinical outcomes and reduce societal burden.[19,20]

In our analysis by outbreak size, we found that outbreaks reported to be of widespread size (country-wide, island nation-wide, or both) were most likely to be detected using our methods (69% sensitivity), and overall, 83% of our start dates were earlier than the issuance dates of matching outbreak reports. Analyzing at a country level, for the 42 countries with reported outbreaks (or in GBD regions with reported outbreaks), we found a similar mean overall sensitivity, but also that it varied by country, with favorable sensitivity and PPV values (of 1.0) for more than half of the countries, but with poor values for a smaller portion of countries. Sensitivity, specificity, PPV, and NPV tended to correlate within GBD regions, with best results commonly found in countries from the Oceania, Caribbean, and Western Sub-Saharan Africa regions (Supplements III and IV, Table S6, and Fig S4, available at www.aaojournal.org).

Geographical spread of conjunctivitis has been reported.[19,20,38] Of note, in some cases, we observed evidence of what seems to be conjunctivitis outbreaks spreading between neighboring countries. For example, this was seen for Haiti and then the Dominican Republic and other nearby countries in the Caribbean GBD region in 2017, as well as for Burkina Faso and then Nigeria and other nearby countries in the same West African GBD region in fall 2016 (Figs 1 and 2; Tables 1 and 2; and Supplemental Table S6 and Fig S4). Some neighboring countries that seem to have been part of the same epidemic were identified only by candidates. For example, see Benin compared with the Dominican Republic in Figure 1 and Supplement IV, Figure S4. This suggests that reports and candidates in 1 country of a GBD may inform increased likelihood of recent or pending outbreaks in neighboring similar

countries, including for those in which there are not sufficient search data. In this respect, among the above-mentioned 6 reported epidemics that we did not include in our primary analysis because of insufficient conjunctivitis-related search data, within their respective GBD regions, corresponding candidate epidemics in the same continuum periods in nearby neighboring countries occurred in 5 of them (Angola, Bonaire, Kiribati, Marshall Islands, and Turks and Caicos Islands). However, it may be difficult to tell whether candidates in neighboring countries result from simultaneous cross-border epidemics or from imprecise geolocation of searches.

However, a number of reported epidemics, especially those categorized as less widespread, were not detected using our approach. In some cases, this was despite sufficient Google search data. Because we analyzed only country-level search data, we may have missed candidates that in future studies may be detectable using search data from smaller regions such as individual states in the United States. The locations where our epidemic detection was the least effective (and scored the lowest) also included countries where Latin and West Germanic languages are less common and where we may have failed to include proper search terms for those countries (such as Eastern Europe, portions of Africa, and the Middle East, where few candidates or reports were found). In addition, we found that for 80 countries, there had been no sufficient search data for any conjunctivitis-related terms. Many of those countries (for example, Azerbaijan and other countries from Central Asia and a large number of Sub-Saharan African countries, including Djibouti and Angola) are in regions where one might expect other languages to be more common (i.e., we did not capture the search language). Some also were very small countries, for example, 14 of 23 countries from Oceania (e.g., Niue, Kiribati) with potentially not enough online users for sufficient search data. Strategies to find additional more appropriate and regionally specific search terms (e.g., *red eye* in world regions where it is used mostly in reference to conjunctivitis), data from other search engines used more often for those regions (such as China or South Korea),[39] or additional signal through inclusion of common search term misspellings may improve the ability to detect candidates in those regions.

A significant fraction of our candidates that did not have matching corresponding outbreak reports (i.e., that could be called false-positive results) were from larger countries. Our comparisons of conjunctivitis candidates with negative control term candidates showed significantly different timing and no evidence of correlation between these 2 groups, implying that conjunctivitis candidates are unlikely because of nonspecific search volume changes (Supplement II, Fig S3). Some reported outbreaks also simply may not have been included in our comparison because our structured approach may not have identified all reports. As an example, Seychelles and Madagascar were not in our originally identified (using a priori queries) corpus of reported outbreaks, and in our analysis, we identified 1 false-positive candidate for each of these countries in the spring of 2015, both of which lowered our sensitivity results. However, a more in-depth search of our cited outbreak report for Réunion[19] revealed that outbreaks did indeed occur in Seychelles and Madagascar and within the same continuum periods as our candidates in those countries, a finding that would have improved our sensitivity and specificity results overall with resulting values of 1.0 for each of those countries (Supplement IV, Fig S4). However, we note that some candidate epidemics may well be spurious, for example, because of spikes in interest when celebrities have conjunctivitis. In some cases, they also could represent another disease in which conjunctivitis is a symptom (e.g., Zika). Our analysis is correlational and investigating social media post content or other online sources of information during candidate epidemic periods may help to determine the reason for searches, thereby improving specificity.

Early awareness, allowing preventative public health responses, can reduce the impact of epidemics. Future improvements of methods such as those presented herein applied prospectively to leverage nontraditional sources of eye health information show promise in providing public health agencies a complementary and relatively low-cost means of improved detection, confirmation, or notification of eye health epidemics.

References

1. Brownstein JS, Freifeld CC. HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveillance*. 2007;12(11):E071129.5.
2. Hartley DM, Nelson NP, Arthur RR, et al. An overview of internet biosurveillance. *Clin Microbiol Infect*. 2013;19(11):1006−1013.
3. Velasco E, Agheneza T, Denecke K, et al. Social media and internet-based data in global systems for public health surveillance: a systematic review. *The Milbank Quarterly*. 2014;92(1):7−33.
4. 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.
5. Brownstein JS, Mandl KD. Reengineering real time outbreak detection systems for influenza epidemic monitoring. *AMIA Symposium*. 2006;2006:866.
6. Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012−1014.
7. 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.
8. Generous N, Fairchild G, Deshpande A, et al. Global disease monitoring and forecasting with Wikipedia. *PLoS Comput Biol*. 2014;10(11):e1003892.
9. Santillana M, Nguyen AT, Dredze M, et al. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol*. 2015;11(10):e1004513.
10. Hoen AG, Keller M, Verma AD, et al. Electronic event-based surveillance for monitoring dengue, Latin America. *Emerg Infect Dis*. 2012;18(7):1147−1150.
11. Allen C, Tsou MH, Aslam A, et al. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. *PloS One*. 2016;11(7):e0157734.

12. Roche B, Gaillard B, Léger L, et al. An ecological and digital epidemiology analysis on the role of human behavior on the 2014 chikungunya outbreak in Martinique. *Sci Rep*. 2017;7(1):5967.
13. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. *PLoS Negl Trop Dis*. 2017;11(1):e0005295.
14. Choi J, Cho Y, Shim E, Woo H. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health*. 2016;16(1):1238.
15. Marques-Toledo CA, Degener CM, Vinhal L, et al. Dengue prediction by the web: tweets are a useful tool for estimating and forecasting dengue at country and city level. *PLoS Negl Trop Dis*. 2017;11(7):e0005729.
16. 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 Ophthalmol*. 2017;135(4):402—403.
17. Smith AF, Waycaster C. Estimate of the direct and indirect annual cost of bacterial conjunctivitis in the United States. *BMC Ophthalmol*. 2009;9:13.
18. Benzekri R, Belfort Jr R, Ventura CV, et al. Manifestations oculaires du virus Zika: où en sommes-nous? *Journal Français d'Ophthalmologie*. 2017;40(2):128—145.
19. Filleul L, Pages F, Wan GC, et al. Costs of conjunctivitis outbreak, Réunion Island, France. *Emerg Infect Dis*. 2018;24(1):168—170.
20. Yen MY, Wu TS, Chiu AW, et al. Taipei's use of a multichannel mass risk communication program to rapidly reverse an epidemic of highly communicable disease. *PLoS One*. 2009;4(11):e7962.
21. Leffler CT, Davenport B, Chan D. Frequency and seasonal variation of ophthalmology-related internet searches. *Can J Ophthalmol*. 2010;45(3):274—279.
22. 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.
23. Deiner MS, Lietman TM, McLeod SD, et al. Surveillance tools emerging from search engines and social media data for determining eye disease patterns. *JAMA Ophthalmol*. 2016;134(9):1024—1030.
24. 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(2):910—920.
25. Institute for Health Metrics and Evaluation. Global burden of diseases regions. http://www.healthdata.org/gbd/faq#What%20countries%20are%20in%20each%20region?. Accessed May 30, 2019.
26. Stocking G, Matsa KE. Using Google Trends data for research? Here are 6 questions to ask. https://medium.com/@pewresearch/using-google-trends-data-for-research-here-are-6-questions-to-ask-a7097f5fb526; 2017. Accessed 25.11.17.
27. Berlinberg EJ, Deiner MS, Porco TC, Acharya NR. Monitoring interest in Herpes Zoster vaccination: analysis of Google search data. *JMIR Public Health Surveill*. 2018;4(2):e10180.
28. Kulldorff M. A spatial scan statistic. *Commun Stat Theory Methods*. 1997;26(6):1481—1496.
29. Wood SN. *Generalized additive models: an introduction with R*. second edition. Boca Raton, FL: CRC Press; 2017.
30. Sie A, Diarra A, Millogo O, et al. Seasonal and temporal trends in childhood conjunctivitis in Burkina Faso. *Am J Trop Med Hyg*. 2018;99(1):229—232.
31. International Society for Infectious Diseases. About ProMED-mail. https://www.promedmail.org/aboutus/. Accessed 17.08.17.
32. Madoff LC. ProMED-mail: an early warning system for emerging diseases. *Clin Infect Dis*. 2004;39(2):227—232.
33. Madoff LC, Woodall JP. The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. *Arch Med Res*. 2005;36(6):724—730.
34. Hossain L, Kam D, Kong F, et al. Social media in Ebola outbreak. *Epidemiol Infect*. 2016;144(10):2136—2143.
35. Institute for Health Metrics and Evaluation. IHME global burden of diseases 2016 location hierarchy file. http://www.healthdata.org/sites/default/files/files/Projects/GBD/IHME_GBD_2016_CODEBOOK.zip. Accessed 1.09.18.
36. Funk S, Gilad E, Watkins C, Jansen VA. The spread of awareness and its impact on epidemic outbreaks. *Proc Natl Acad Sci U S A*. 2009;106(16):6872—6877.
37. Rashid H, Ridda I, King C, et al. Evidence compendium and advice on social distancing and other related measures for response to an influenza pandemic. *Paediatr Respir Rev*. 2015;16(2):119—126.
38. Jawetz E. The story of shipyard eye. *Br Med J*. 1959;1(5126):873—876.
39. Search engine market share by country: 2015 update. https://returnonnow.com/internet-marketing-resources/2015-search-engine-market-share-by-country/; 2015. Accessed 15.01.19.

Footnotes and Financial Disclosures

Originally received: August 29, 2018.
Final revision: March 15, 2019.
Accepted: April 5, 2019.
Available online: April 11, 2019. Manuscript no. 2018-1973.

[1] F. I. Proctor Foundation, University of California, San Francisco, San Francisco, California.

[2] Department of Ophthalmology, University of California, San Francisco, San Francisco, California.

[3] Department of Ophthalmology, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts.

[4] Department of Epidemiology and Biostatistics, Global Health Sciences, University of California, San Francisco, San Francisco, California.

Dr. Stephen D. McLeod, the editor-in-chief of the journal, was recused from the peer-review process of this paper.

Financial Disclosure(s):
The author(s) have made the following disclosure(s): J.C.: Consultant — Shire.

Supported in part by the National Eye Institute, National Institutes of Health, Bethesda, Maryland (grant nos.: 1R01EY024608-01A1 [T.M.L., T.C.P., M.S.D., S.D.M., J.C.] and EY002162 [M.S.D., S.D.M., J.W., T.M.L., T.C.P.]); and Research to Prevent Blindness, Inc., New York, New York (unrestricted grant [M.S.D., S.D.M., J.W., T.M.L., T.C.P.]). The sponsor or funding organization had no role in the design or conduct of this research.

HUMAN SUBJECTS: No human subjects were included in this study. The human ethics committees at the University of California, San Francisco, approved the study. All research adhered to the tenets of the Declaration of Helsinki. The requirement for informed consent was waived because of the retrospective nature of the study.

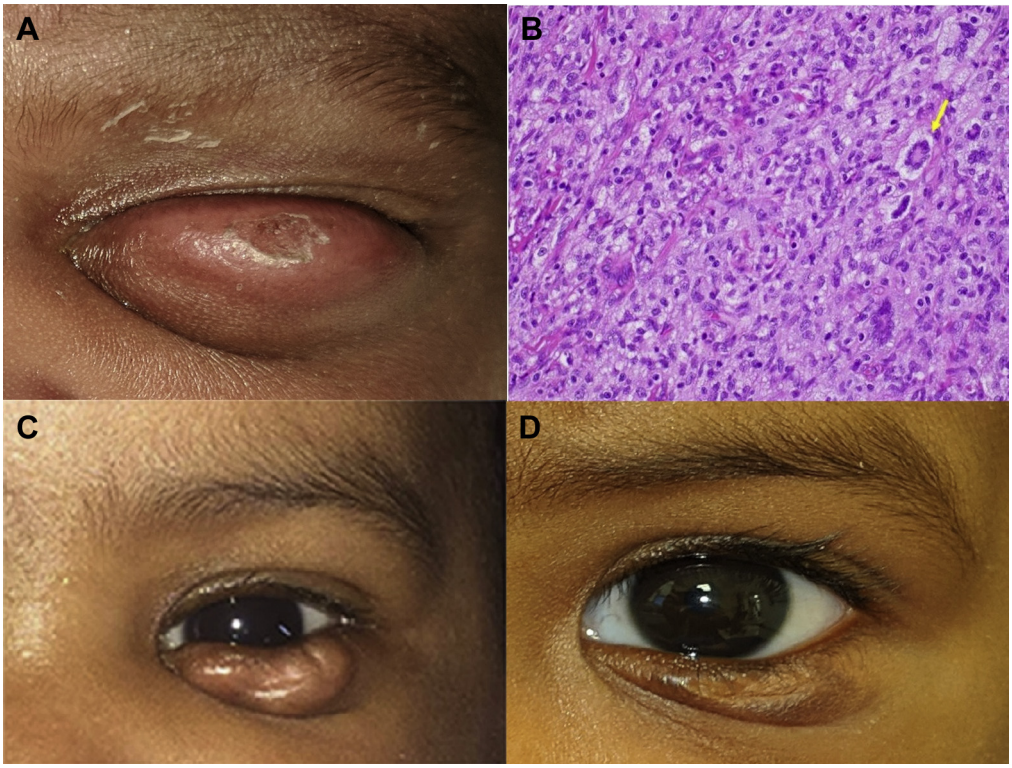
No animal subjects were included in this study.

Author Contributions:
Conception and design: Deiner, McLeod, Wong, Chodosh, Lietman, Porco
Analysis and interpretation: Deiner, McLeod, Wong, Lietman, Porco
Data collection: Deiner, Lietman, Porco
Obtained funding: Lietman
Overall responsibility: Deiner, McLeod, Chodosh, Lietman, Porco

Abbreviations and Acronyms:
CI = confidence interval; **GBD** = Global Burden of Diseases; **NPV** = negative predictive value; **PPV** = positive predictive value; **ProMED** = Program for Monitoring Emerging Diseases; **SD** = standard deviation.

Correspondence:
Travis C. Porco, PhD, MPH, F. I. Proctor Foundation, University of California, San Francisco, 513 Parnassus, San Francisco, CA 94143. E-mail: travis.porco@ucsf.edu.

Pictures & Perspectives



Congenital Macronodular Eyelid Juvenile Xanthogranuloma

A 2-week-old boy presented with a congenital left eyelid mass. Examination disclosed a large, firm left lower eyelid nodule spanning the width of the lid and causing closure of the palpebral fissure (Fig **A**). There did not appear to be postseptal involvement. Incisional biopsy revealed a dense lymphohistiocytic infiltrate with scattered Touton giant cells (Fig **B**, *yellow arrow*) consistent with juvenile xanthogranuloma (JXG). A single intralesional dexamethasone injection was performed and the child displayed good improvement during follow-up (Fig **C**, 4 weeks; Fig **D**, 11 months). Congenital eyelid JXG is quite rare and can be treated with intralesional steroid. (Magnified version of Fig **A**-**D** is available online at www.aaojournal.org).

Adam Chubak, MD[1]
Mohamed Kahila, MD[2]
Roman Shinder, MD, FACS[1]

[1]*Ophthalmology, SUNY Downstate Medical Center, Brooklyn, New York;* [2]*Pathology, Kings County Hospital Center, Brooklyn, New York*