Title
Google Searches and Detection of Conjunctivitis Epidemics Worldwide

Journal
Ophthalmology, 126(9)

ISSN
0161-6420

Authors
Deiner, Michael S
McLeod, Stephen D
Wong, Jessica
et al.

Publication Date
2019-09-01

DOI
10.1016/j.ophtha.2019.04.008

Peer reviewed

HHS Public Access

Author manuscript

Ophthalmology. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

Ophthalmology. 2019 September ; 126(9): 1219–1229. doi:10.1016/j.ophtha.2019.04.008.

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]

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

[2]Department of Ophthalmology, University of California San Francisco, San Francisco, CA, USA

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

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

Abstract

Purpose and Objective: Epidemic and seasonal infectious conjunctivitis outbreaks can adversely impact education, workforce and economy. Yet conjunctivitis is typically 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 were then 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 time period by our candidate epidemics, identified from conjunctivitis-related searches

Results: We identified 135 candidate conjunctivitis epidemic periods from 77 countries. Compared to our *a priori* defined collection of known reported outbreaks, candidate conjunctivitis epidemics identified 18 out of 26 (69% sensitivity) of the reported country-wide and/or island

Corresponding Author and Address for Reprints: Travis C. Porco, F.I. Proctor Foundation, University of California San Francisco, 513 Parnassus, San Francisco, CA 94143 (travis.porco@ucsf.edu).

Conflict of Interest:
No conflicting relationship exists for any author.

nation-wide outbreaks, 9 out of 20 (45% sensitivity) of the reported region and/or district-wide outbreaks, but far fewer nosocomial and reported smaller outbreaks. Similar overall and individual sensitivity, as well as specificity, was found on a country-level basis. We also found that 83% of our candidate epidemics had start dates prior (of those, 20% were over 12 weeks prior) to 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 would suggest (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 over 80% detected prior to 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 further improve sensitivity and cross-validate detected candidates, allowing identification of candidate conjunctivitis epidemics from Internet search data to potentially complementarily benefit traditional reporting and detection systems to improve epidemic awareness.

Precis

Although burdensome, conjunctivitis is typically not reportable. In this masked study, we found Google Trends search data analysis can identify known conjunctivitis outbreaks worldwide, suggesting potential future benefit for public health agencies monitoring eye disease.

Introduction

"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 which may be underreported or not generally reportable. These include conjunctivitis[16], a condition only reportable in the USA 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].

In the past, evidence has been found that age- and etiology-specific features of conjunctivitis clinical epidemiology and seasonal patterns were partially reflected in available online search and social media data sources, including tweets, blogs posts, forums and search engine query data[21–24]. Here, we tested the hypothesis that Internet search data for keywords relevant to conjunctivitis could be used to identify actual epidemics of conjunctivitis. 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 we validated our detection method using those reports.

Google search data for identifying epidemics

The Google Timeline for Health API (application programming interface) allows researchers and others, after applying to Google and being granted permission, to access Google Trends data about geo-temporal location of online searches and has been used for research for example to study behavior, explore outbreaks and forecast economic activity[26,27]. Using this API we collected worldwide national-level daily Google search data for 24 keywords. These keywords included terms related to conjunctivitis in several languages, as well as positive control keywords related to other diseases, and negative control keywords designed to reflect general changes in search data volume (to account for any search volume changes presumably unrelated to disease).

For a list of keywords, please see Supplement VI. Data were obtained for the period July 10, 2012 to July 9, 2017. The resulting time series represents relative search interest data, reflecting the proportion of searches for the term of interest out of all searches for all terms for a given geography and time period. The proportion is calculated by Google using a random sample. Very small values are censored by the API, partly to protect privacy; such censored values appear as zeros in the time series. Country-search term combinations yielding no relative search information were excluded.

Identification of candidate epidemics

To retrospectively identify epidemics, we used an approach that implemented three 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 due to media coverage. It then examined a 31-day centered moving average. The second algorithm (“Lush”) was then automatically implemented but only for 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 was then 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, which 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 was then 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 was thus accomplished by examining candidates from this automatically applied collection of algorithms. Further details are provided in Supplement V. For each epidemic, the first date at which the threshold was exceeded was considered the earliest detectable date, i.e., 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 was also not used. After all candidate epidemics were identified, masked terms and geolocation were unmasked for subsequent validation comparisons to reported outbreaks.

Conjunctivitis outbreak reports: ProMED, PubMed, and Other Online

Our study is designed to identify conjunctivitis outbreaks which, in many countries, are not reported in any standardized or systematic manner. Although a true gold standard is therefore not available, other sources of reports can serve to validate our candidate epidemics. In late summer of 2017, we identified conjunctivitis outbreak reports for the period from July 10, 2012 through August 9, 2017 using 3 sources: ProMED, PubMed, and Other Online Internet content. ProMED (the Program for Monitoring Emerging Diseases) 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. ProMED has been an early warning system for infectious diseases for over 22 years[31–34]. We queried 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 three 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 and/or island nation-wide”, “region and/or district wide”, “nosocomial”, or “small” (e.g., one classroom, but not associated with a health care facility). We excluded reports with unclear start dates (less precise than a one month time window) or start dates not occurring between July 18, 2012 and July 2, 2017. We excluded one 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, multiple candidate epidemics identified from analysis of Google search data may be close together in time, for a given outbreak. To compare three report sources and identified candidates for each outbreak, we identified one 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 or more days separated consecutive reports in the same country, we considered a new epidemic (new continuum ID 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 two ways. First, for each reported outbreak in a country, we determined whether or not at least one candidate epidemic was identified within the same continuum. Second, we repeated the analysis for the four 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 to conjunctivitis report dates. Confidence intervals for proportions were computed using the exact (Clopper-Pearson) method.

Of course, in the absence of a gold standard, the sensitivity, specificity, and positive and negative predictive values (PPV, NPV) of our candidate epidemics cannot, *sensu stricto,* be computed. However, on a country- based level, we estimated a measure of these quantities based on the following assumptions: (a) The 40 total possible 45-day continuum periods (1825 days/45 days) in the time series for a country serves as an effective denominator of 40 (a max possible of 40 continuum periods per country, that each could have had a candidate in them) (b) any continuum period containing a start date for a reported outbreak is considered to reflect a true epidemic, and any not containing one is assumed to have experienced no epidemic and (c) any continuum period containing a start date for a candidate conjunctivitis epidemic effectively "tests positive", while any continuum period not containing a candidate corresponds to a negative test. All results were first calculated for countries containing reported outbreaks. For countries in which there was no reported outbreak at all, but for which we found candidate epidemics, countries were then grouped into similar regions, defined using the IHME GBD 2016 location hierarchy file as a guide[25,35] to assess if "false positives" (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 only assessed for specificity and NPV. The mean sensitivity and other summary values of all the country-level results were also computed as a secondary comparison to our other all-countries analysis.

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 vs. 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 to 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 country at random, and within countries, conducted a random cyclic permutation of the starting times (accounting for the non-independent 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 non-epidemic 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 five year period as being part of a candidate epidemic or not. Using these binary series, we compared the phi correlation based on conjunctivitis search terms and search terms related to influenza, allergy and negative control ("for" and "para") terms.

IRB Approval

UCSF IRB Approval was obtained prior to this study (approval# 14-14743).

Results

Outbreak Reports

A priori, before any comparison to candidate epidemics, we identified 87 conjunctivitis outbreak reports from 49 countries from July 17, 2012 and July 2, 2017. We excluded reports from countries yielding no Google search information (3 ProMED, 1 PubMed, and 4 Other Online outbreak reports—please 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 one 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 to our candidate epidemics, as shown in 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 and/or scan methods were often 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), "conjuntivitis" (n=26, 19% of total). For some conjunctivitis search terms specific to certain locations, we often only detected 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, for term "apollo eye" were found 4 times and only in Nigeria, and for term "azoumounou" (Haitian creole) were found only for Haiti (3 times) and USA (1 time). More details of the results for all search terms can be seen in Supplement I Tables S3-S5.

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

Detected candidate epidemics 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. This table also allows a comparison of day differences between candidate conjunctivitis epidemic start dates (the leading edge of the scan window for which the epidemic threshold was first reached) and the start date of each the reported outbreak in the same 45 day continuums, and allows a comparison with the report's issuance date. More detailed analysis of these results, including sensitivity, start or report date differences, frequently validated keywords, and percentage of candidates validated by reports is described below (as well as in the Supplement). Daily searches, detected candidates and outbreak reports are visually compared in time series Figures 1-2. 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 candidates epidemics identified are shown as red triangles with their identified start date, the top row shows 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 a 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 ProMED, PubMed and Online Other outbreak reports

Our method identified 28 out of 56 (50% sensitivity, 95% CI: 36% to 63%) of the reported outbreaks (see Figure 2, Tables 1-2). We identified 18 out of 26 (69% sensitivity, 95% CI: 48% to 86%) of the reported country-wide and/or island nation-wide outbreaks, 9 out of 20 (45% sensitivity) of the reported region and/or district-wide, 1 out of 4 (25% sensitivity) of the reported nosocomial, and 0 out of 6 (0% sensitivity) of the 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 upon 31 or 60 day continuum periods, and found no sensitivity differences from that reported above for when using a 45-day period. Second, using an alternative time window based approach to assess sensitivity we found similar sensitivity results as with the approaches described above (50% overall, 95% CI: 37% to 63%; by size: 68% country-wide and/or island nation-wide; 45% region and/or district-wide; 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% to 17%) when comparing to the reported outbreaks.

Validations by Country

We also analyzed results on a country level. Please see Supplement III Table S6 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, sd: 0.03, min: 0.81, max: 1),

mean NPV per country was 0.995 (median: 1, sd: 0.014, min:0.91, max: 1), and the mean number of false positive continuums per country was 0.67 (median: 0, sd: 1.18, min: 0, max: 7). For just the 42 countries with any reported outbreaks specificity, NPV and mean number of false positives 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, min: 0, max: 1) and mean adjusted (assigning 0 if no candidates were found) PPV was 0.55. For countries with no reported outbreaks, the mean number of false positives continuums per country was 0.64.When grouping countries by GBD region, we also found that many countries with no gold standard to compare to had a reported outbreak in a neighboring country within their continuum period within that GBD region (See examples in Supplement IV Figure S4). Therefore, for our analysis and in the table we have 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 improves to a mean of 0.49 continuums (out of a maximum of 40 possible continuums) per country.

Start date comparisons

We compared our candidate epidemic start dates to 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 prior to their matching reported outbreak start date). A total of 11% of the candidate start dates were 1 to 3 weeks prior to their matching outbreak report's reported start date, and 11% were 4 to 6 weeks prior.

Report issuance date comparisons

When comparing our candidate epidemic start dates to 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 prior to their matching reported epidemic's report issuance date). Of those, 20% of our candidate start dates were 1 to 3 weeks prior to their matching report's issuance date, 17% were 4 to 6 weeks prior, 9% were 7 to 12 weeks prior and 20% were over 12 weeks prior.

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 which are closer to reported conjunctivitis outbreaks than chance alone ($P = 0.40$). A measure of specificity was also computed by determining the fraction of non-epidemic days which are not 31 days after an identified candidate epidemic. For conjunctivitis candidates this value was 95.4% (95% CI: 94.0% to 96.2%). A similar result was found for negative control term candidates.

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

Discussion

Overall, our study has 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 better identify and take action in response to outbreaks, including for conjunctivitis[19]. Although the issuance date of some formal public health agency reports presumably can lag simply due to administrative delays, in some cases delays may be due to 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 accelerate awareness, confirmation, and official public health reporting. For infectious epidemics, studies suggest there is a benefit of reducing the impact of outbreaks[36] including through social distancing reducing transmission, such as for flu[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 and/or island nation-wide) 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 that it varied by country, with favorable sensitivity and PPV values (of 1.0) for over half of the countries, but on the other hand 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 (see Supplement III-IV Table S6 and Figure S4).

Geographical spread of conjunctivitis has been reported[19,20,38] Of note, in some cases, we observed evidence of what appears to be conjunctivitis outbreaks spreading between neighboring countries. For example, this was seen for Haiti and then 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 the fall of 2016 (please see Figures 1, 2, Tables 1-2 and Supplement Table S6 and Figure S4). Some neighboring countries apparently part of the same epidemic only were identified by candidates, for example see Benin compared to Dominican Republic in Figure 1 and

Supplement IV Figure S4. This suggests reports and candidates in one country of a GBD may inform increased likelihood of recent or pending outbreaks in neighboring similar countries, including for those in which there is not sufficient search data. In this respect, for the previously above-mentioned 6 reported epidemics that we did not include in our primary analysis due to insufficient conjunctivitis-related search data, within their respective GBD regions, five of them (Angola, Bonaire, Kiribati, Marshall Islands, Turks and Caicos Islands) had corresponding candidate epidemics in the same continuum periods in nearby neighboring countries. However, it may be difficult to tell whether candidates in neighboring countries result from simultaneous cross-border epidemics or from imprecise geolocation of searches.

A number of reported epidemics though, especially those categorized as less widespread, were not detected using our approach. In some cases this was despite sufficient Google search data. As we only analyzed 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 USA 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 the 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 were also 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), or 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 might improve 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 positives”) were from larger countries. Our comparisons of conjunctivitis candidates to negative control term candidates showed significantly different timing and no evidence of correlation between these two groups, implying conjunctivitis candidates are unlikely due to non-specific search volume changes though (see Supplement II Figure S3). Some reported outbreaks may also simply not have been included in our comparison, since 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 one “false positive” candidate for each of these countries in the spring of 2015, both of which lowered our sensitivity results. A more in-depth search of our cited outbreak report for Réunion[19] however, reveals outbreaks did indeed occur in Seychelles and Madagascar and within the same continuum periods as our candidates in those countries, a finding which would have improved our sensitivity and specificity results overall with resulting values of

1.0 for each of those countries (see Supplement IV Figure S4). We note that some candidate epidemics may well be spurious though, for example due to spikes in interest when celebrities have conjunctivitis. In some cases they could also represent another disease where conjunctivitis is a symptom (e.g. Zika). Our analysis is correlational, and investigating social media post contents, or other online sources of information, during candidate epidemic periods may help determine the reason for searches and thereby improve specificity.

Early awareness, allowing preventative public health responses, can reduce the impact of epidemics. Future improvements of methods such as those presented here applied prospectively to leverage non-traditional 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding/Support

This work was supported in part by grant 1R01EY024608-01A1 (Lietman, PI) from the National Institutes of Health National Eye Institute (NIH-NEI), grant EY002162 (Core Grant for Vision Research -Ullian, PI) from the NIH-NEI, and an Unrestricted Grant from Research to Prevent Blindness (McLeod, PI). The sponsor or funding organization had no role in the design or conduct of this research.

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. Clinical microbiology and infection. 2013;19(11): 1006–1013. doi:10.1111/1469-0691.12273 [PubMed: 23789639]
3. 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. The Milbank Quarterly. 2014;92(1):7–33. doi: 10.1111/1468-0009.12038 [PubMed: 24597553]
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. doi:10.1371/journal.pone.0109583 [PubMed: 25337815]
5. Brownstein JS, Mandl KD. Reengineering real time outbreak detection systems for influenza epidemic monitoring. AMIA Symposium. 2006:866. [PubMed: 17238486]
6. 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. doi:10.1038/nature07634 [PubMed: 19020500]
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. doi:10.1371/journal.pone.0090536 [PubMed: 24599062]
8. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with wikipedia. PLoS Computational Biology. 2014;10(11):e1003892. doi:10.1371/journal.pcbi.1003892 [PubMed: 25392913]

9. 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 Computational Biology. 2015;11(10):e1004513. doi:10.1371/journal.pcbi.1004513 [PubMed: 26513245]
10. Hoen AG, Keller M, Verma AD, Buckeridge DL, Brownstein JS. Electronic event-based surveillance for monitoring dengue, Latin America. Emerging infectious diseases. 2012;18(7):1147–1150. doi:10.3201/eid1807.120055 [PubMed: 22709430]
11. Allen C, Tsou MH, Aslam A, Nagel A, Gawron JM. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. PloS one. 2016;11(7):e0157734. doi:10.1371/journal.pone.0157734 [PubMed: 27455108]
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. Scientific reports. 2017;7(1):5967. doi:10.1038/s41598-017-05957-y [PubMed: 28729711]
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 neglected tropical diseases. 2017;11(1):e0005295. doi:10.1371/journal.pntd.0005295 [PubMed: 28085877]
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. [PubMed: 27931204]
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. [PubMed: 28719659]
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 Ophthalmology. 2017;135(4):402–403. doi:10.1001/jamaophthalmol.2017.0140
17. Smith AF, Waycaster C. Estimate of the direct and indirect annual cost of bacterial conjunctivitis in the United States. BMC ophthalmology. 2009;9:13. doi:10.1186/1471-2415-9-13 [PubMed: 19939250]
18. Benzekri R, Belfort R Jr., Ventura CV, et al. Manifestations oculaires du virus Zika: Oùen sommes-nous? Journal Français d'Ophthalmologie. 2017;40(2):128–145.
19. Filleul L, Pages F, Wan GC, Brottet E, Vilain P. Costs of Conjunctivitis Outbreak, RÃ©union Island, France. Emerging Infect Dis. 2018;24(1):168–170. [PubMed: 29260662]
20. Yen MY, Wu TS, Chiu AW, et al. Taipei's use of a multi-channel mass risk communication program to rapidly reverse an epidemic of highly communicable disease. PLoS ONE. 2009;4(11):e7962. [PubMed: 19956722]
21. Leffler CT, Davenport B, Chan D. Frequency and seasonal variation of ophthalmology-related internet searches. Canadian journal of ophthalmology. 2010;45(3):274–279. doi:10.3129/i10-022 [PubMed: 20436544]
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. doi:10.1111/all.12528 [PubMed: 25280183]
23. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance tools emerging from search engines and social media data for determining eye disease patterns. JAMA Ophthalmology. 2016;134(9):1024–1030. doi:10.1001/jamaophthalmol.2016.2267 [PubMed: 27416554]
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. [PubMed: 29450538]
25. Global burden of diseases regions. http://www.healthdata.org/gbd/faq#What%20countries%20are%20in%20each%20region? Accessed September 1, 2018.
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. Accessed November 25, 2017.
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. [PubMed: 29720364]

28. Kulldorff M A spatial scan statistic. Communications in Statistics - Theory and Methods. 1997;26(6):1481–1496. doi:10.1080/03610929708831995
29. Wood SN. Generalized Additive Models: An Introduction with R, Second Edition. 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. [PubMed: 29761759]
31. ProMED-mail. https://www.promedmail.org/aboutus/. Accessed August 17, 2017.
32. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis. 2004;39(2):227–232. [PubMed: 15307032]
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. [PubMed: 16216654]
34. Hossain L, Kam D, Kong F, Wigand RT, Bossomaier T. Social media in Ebola outbreak. Epidemiol Infect. 2016;144(10):2136–2143. [PubMed: 26939535]
35. 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 September 1, 2018.
36. Funk S, Gilad E, Watkins C, Jansen VA. The spread of awareness and its impact on epidemic outbreaks. Proc Natl Acad Sci USA. 2009;106(16):6872–6877. [PubMed: 19332788]
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. PaediatrRespirRev. 2015;16(2):119–126.
38. Jawetz E The story of shipyard eye. Br Med J. 1959;1(5126):873–876. [PubMed: 13629151]
39. Search engine market share by country: 2015 update. https://returnonnow.com/internet-marketing-resources/2015-search-engine-market-share-by-country/. Accessed January 15, 2019.

Figure 1. Five illustrative countries demonstrating daily search data, candidate epidemics identified from that data, and reported outbreaks

For each country (column), the timespan provided is from the earliest to latest occurring candidate conjunctivitis epidemics or reported epidemics, within the full study period (i.e. first four 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 five countries had no available daily relative search interest values. (Legend: Y- axis for each time series is normalized (% of max value) daily search values. Daily values are indicated by colored vertical bars; ProMED, 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. Time series of reported outbreaks compared to 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 and/or candidates occurring within the same continuum period. For all reported outbreaks that had an issuance (publication or first online) date that was one or more weeks 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 ID periods and compare to candidates) were much earlier than when the report was actually issued (e.g. see Réunion, Tonga). Countries with no reported outbreaks, not shown. (Legend: Gold inverted triangles represent issuance date of ProMED, 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).

Table 1.

Country-Wide and Island Nation-Wide outbreak reports, compared to identified candidate epidemics.

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 to dates and locations of identified candidate epidemics (note, only one report per continuum, that with the earliest reported start date, was used in calculating sensitivity in this study). Columns: “Country” represents name of the country; “Found” indicates whether if candidate was found, in a masked fashion, within the same 45 day continuum as the reported outbreak; “Report Source” indicates the source from where outbreak reports were obtained using queries of ProMED, PubMed, and Other Online (Other) reports; “Reported Start” indicates the start date of the outbreak defined in the report; “Days Prior to Start” indicates 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); “Report Issuance” indicates the date the report was published; “Days Prior to Issuance” indicates 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); “Reference” indicates the cited original source of the reported outbreak. Please see Supplement 1 for Table 1 Outbreak Report references.

Country	Found	Report Source	Reported Start	Days Prior to Start	Report Issuance	Days Prior to 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

Country	Found	Report Source	Reported Start	Days Prior to Start	Report Issuance	Days Prior to Issuance	Report Reference
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
Viet Nam	Yes	Other	2013-09-01	−15	2013-09-20	4	1z

Table 2.
Smaller reported outbreaks, compared to identified candidate epidemics.

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) are included in Table 2. As with Table 1, 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. Table 2 column descriptions and comparisons to candidates identified from Google search data are also all as described for Table 1. Please see Supplement 1 for Table 2 Outbreak Report references.

Country	Found	Report Source	Reported Start	Days Prior to Start	Report Issuance	Days Prior to Issuance	Report Reference
Outbreak Size Group -District/Region-Wide:							
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
Outbreak Size Group -Nosocomial:							
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
Outbreak Size Group -Small:							
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

Country	Found	Report Source	Reported Start	Days Prior to Start	Report Issuance	Days Prior to Issuance	Report Reference
United States	No	Other	2016-07-20		2016-07-20		2dd