Tracking Outbreaks of Febrile Infectious Diseases in India Using Google Trends

Pranab Chatterjee
Department of Community Medicine, University College of Medical Sciences

Abstract

Background: Search engine activity has been used as an indicator of disease incidence in developed countries. Non-availability of high quality data on Internet use and disease surveillance has hampered similar studies in developing countries.

Objective: To evaluate the correlation between search volumes for “fever” on Google and febrile infectious disease outbreaks reported by the Integrated Disease Surveillance Program.

Methods: Data on Google search volume for “fever” between January 2014 and December 2014 from India was downloaded from the Google Trends Insights website. Weekly data on outbreaks and case counts of infectious causes of febrile illnesses (dengue, PUO/fever, chikungunya, and typhoid) were obtained from the Integrated Disease Surveillance Program website. Spearman’s rho was calculated to estimate unadjusted correlation between search activities with weekly disease metrics. Time-series analysis of Google search query volume and disease metrics was done to ascertain whether they shared a common stochastic drift using the two-step Engle-Granger method.

Results: The unadjusted correlation was statistically significant between search activity and dengue outbreak (r=0.632, p<0.001) and cases (r=0.673, p<0.001); PUO/fever outbreaks (r=0.315, r=0.026) and cases (r=0.323, p=0.022); chikungunya cases (r=0.374, r=0.007); and total outbreaks (r=0.581, r=0.001) and cases (r=0.615, p<0.001). The test for cointegration of search activity showed stationarity of residuals with both disease outbreaks (p=0.048) and case counts (p=0.041).

Conclusions: There was agreement in the search volume with disease outbreaks and case counts. The time trends of search activity were cointegrated with both disease outbreaks and case counts on a weekly basis. This indicated that search activity and infectious diseases outbreaks and case counts were related. However, owing to the non-representative nature of the disease metrics and non-uniform access and limited use of Internet to seek health information, it would be inappropriate to use search engine query volume to predict or forecast disease outbreaks in the Indian context.

Introduction

An increasing number of people from developed countries have been shown to use Internet search engines to seek health-related information [1] [2]. Naturally, this has led to the question of whether looking at online search volumes would be indicative of disease burdens. It was established by investigators from Google that it was possible to detect influenza epidemics in large areas with a high density of Internet users [3]. Investigators have also established that during the 2009 influenza season, which experienced very high disease activity, it was possible to predict the pandemic H1N1 waves in Manitoba using Google Flu Trends and emergency department triage data [4]. Similar results were replicated in South China [5], New Zealand [6], and thirteen European countries (Belgium, France, Hungary, Netherlands, Norway, Poland, Spain, Sweden, Switzerland, Bulgaria, Germany, Russian Federation, and Ukraine), [7] and in all cases there was good agreement between search engine query volumes and actual, reported disease burden. Gunther Eysenbach used this changing behavior pattern to propose a new research discipline: information epidemiology or infodemiology, which he defined to be “the study of the determinants and distribution of health information and misinformation, which may be useful in guiding health professionals and patients to quality health information on the Internet” [8].

However, not many similar studies have been conducted in the setting of developing countries, where Internet search engine use and health-seeking behavior may be very different as compared to that in developed nations. In a study led by Google investigators, aggregated, anonymized search volumes for terms related to “dengue” were found
to fit well with the actual number of cases of dengue reported from Bolivia, Brazil, India, Indonesia and Singapore [9]. In this study, some discrepancy in the fit was observed in the case of India possibly owing to issues of poor penetration and adoption of Internet in the rural areas.

With these factors in mind, the current study endeavored to look at search volumes for the search term “fever” and the corresponding number of outbreaks and cases of febrile infectious diseases (dengue, PUO (Pyrexia of Unknown Origin)/fever, chikungunya, and typhoid) reported by the Integrated Disease Surveillance Program (IDSP) in India.

**Objective**

The objective of the study was to evaluate the correlation between search volumes for “fever” on Google and febrile infectious disease outbreaks reported by the Integrated Disease Surveillance Program.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>55.1381</td>
<td>3.21775</td>
<td>17.1356</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>Dengue Outbreaks</td>
<td>3.25505</td>
<td>0.760658</td>
<td>4.2793</td>
<td>0.00010</td>
</tr>
<tr>
<td>PUO/Fever Outbreaks</td>
<td>1.1477</td>
<td>0.815971</td>
<td>1.4065</td>
<td>0.16643</td>
</tr>
<tr>
<td>Chikungunya Outbreaks</td>
<td>-1.31275</td>
<td>1.51664</td>
<td>-0.8656</td>
<td>0.39132</td>
</tr>
<tr>
<td>Typhoid Outbreaks</td>
<td>1.33222</td>
<td>2.65835</td>
<td>0.5011</td>
<td>0.61871</td>
</tr>
</tbody>
</table>

a

Unit-root null hypothesis: $a = 1$ Test without constant
Model: $(1-L)y = (a-1)y(-1) + ... + e$

- 1st-order autocorrelation coefficient for $e$ = -0.016
- Lagged differences $F(2, 44) = 3.315 [0.0456]$
- Estimated value of $(a - 1)$ = -0.288812
- Test statistic: tau nc(1) = -1.95427
- Asymptotic p-value = 0.04849

b
<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>58.2016</td>
<td>2.76838</td>
<td>21.0237</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>Dengue Cases</td>
<td>0.0249777</td>
<td>0.00678216</td>
<td>3.6828</td>
<td>0.00062</td>
</tr>
<tr>
<td>PUO/Fever Cases</td>
<td>0.0212442</td>
<td>0.0172872</td>
<td>1.2289</td>
<td>0.22550</td>
</tr>
<tr>
<td>Chikungunya Cases</td>
<td>0.0526435</td>
<td>0.0335959</td>
<td>1.5670</td>
<td>0.12413</td>
</tr>
<tr>
<td>Typhoid Cases</td>
<td>-0.0226018</td>
<td>0.0368249</td>
<td>-0.6138</td>
<td>0.54246</td>
</tr>
</tbody>
</table>

Mean dependent var | 65.80000 | S.D. dependent var | 13.21564 |
Sum squared resid  | 5513.641 | S.E. of regression | 11.06912 |
R-squared           | 0.355733 | Adjusted R-squared | 0.298464 |
F(4, 45)            | 6.211691 | P-value(F)         | 0.000455 |
Log-likelihood      | -188.5209| Akaike criterion   | 387.0417 |
Schwarz criterion   | 396.6018 | Hannan-Quinn       | 390.6823 |
rho                 | 0.571729 | Durbin-Watson      | 0.838395 |

Unit-root null hypothesis: a = 1  Test without constant
Model: (1-L)y = (a-1)*y(-1) + ... + e
1st-order autocorrelation coefficient for e | 0.003
Lagged differences | F(2, 44) = 2.361 [0.1061]
Estimated value of (a - 1) | -0.286492
Test statistic: tau_n=1 | -2.02102
Asymptotic p-value | 0.04149

![Graph showing trends and correlations between different infectious disease cases and Google Trends activity.](image-url)
Figure Legend

Figure A: Result of the OLS step for Model 1 (disease outbreaks and search volume)
Figure B: Result of the Augmented Dickey-Fuller Test on Residuals of Model 1 (disease outbreaks and search volume)
Figure C: Result of the OLS Step for Model 2 (case counts of diseases and search volume)
Figure D: Result of the Augmented Dickey-Fuller Test on Residuals of Model 2 (case counts of diseases and search query volume)
Figure E: Time series trends of Google Search volumes (secondary axis on the right) with number of cases of infectious diseases (left side, primary axis) causing fever per week
Figure F: Time series trends of Google Search volumes (secondary axis on the right) with number of outbreaks of infectious diseases causing fever per week
Figure G: Time series trends of Google Search volumes (right side, secondary axis) with total outbreaks (right side, secondary axis) and total cases (left side, primary axis) per week

Google search query volume from India, between January 5, 2014 and December 14, 2014, was downloaded from the Google Trends Insights website (http://google.co.in/trends). Google Trends data are scaled and normalized using specific algorithms to ensure that data remains comparable despite different denominators (total users using the search engine) across time.

Weekly count of outbreaks and cases of the identified diseases was downloaded from the publicly available reports on the IDSP website (http://idsp.nic.in). The objective of the study was to show that there was a relationship between the time trends of the search query volume and reported disease outbreaks and case counts.

Spearman’s rho was calculated to estimate unadjusted correlation between search activities with weekly disease metrics (counts of outbreaks and case counts). Time-series
analysis of search activity and disease metrics was done to ascertain whether they shared a common stochastic drift. The two-step Engle-Granger method was used to ascertain cointegration under the assumption that the variables were integrated to the order of one. In the first step the dependent variable (Google search query volume) was regressed on a constant and the independent variables (outbreaks of the identified diseases in model 1 and case counts of the identified diseases in model 2) and the residuals were calculated. In the next step, the Augmented Dickey-Fuller test was run on the residuals (without a constant term) under the null hypothesis that the Google search query volume and disease outbreaks (model 1) and case counts (model 2) were not cointegrated. Rejection of the null hypothesis would be evidence that the residual is stationary, that is, the time series are cointegrated, thus proving that the Google search query trend underwent similar changes with changes in the outbreak numbers (model 1) or case counts (model 2) through time, on a week-by-week basis.

Data was entered into MS Excel and statistical tests for time series data were done using the open-source statistical software Gretl version 1.9.92.

Results & Discussion

RESULTS:
The unadjusted correlation was statistically significant between search activity and dengue outbreak ($r=0.632, p<0.001$) and cases ($r=0.673, p<0.001$); PUO/fever outbreaks ($r=0.315, p=0.026$) and cases ($r=0.323, p=0.022$); chikungunya cases ($r=0.374, p=0.007$); and total outbreaks ($r=0.581, p<0.001$) and cases ($r=0.615, p<0.001$).

In Model 1, the dependent variable was the Google search query volume; the independent variables in the model included outbreaks of dengue, PUO/fever, chikungunya, and typhoid. In Model 2, the dependent variable was the Google search query volume; the independent variables in the model included case counts (incidence) of dengue, PUO/fever, chikungunya, and typhoid.

Tables 1 and 3 show the result of the regression done in the first step of the Engle-Granger method for models 1 and 2 respectively. The residuals from each of the models were saved as another variable in Gretl and the results of the Augmented Dickey-Fuller Tests run on that variable for model 1 and 2 separately are shown in Tables 2 and 4 respectively.

Figure 1 shows time series trend between Google search query volume (measured on the secondary axis on the right) and the number of cases of infectious diseases causing fever. Figure 2 shows time series trend between Google search query volume (measured on the secondary axis on the right) and total outbreaks every week. Figure 3 shows the time series trend between Google search query volume and the total numbers of outbreaks and cases per week.

DISCUSSION:
The time series plots show a weak temporal association between the total numbers of outbreaks every week and the total number of cases of different infectious diseases causing fever every week and Google search query volume. The search activity goes up slightly before every spike in disease incidence. However, because the baseline search activity is very high, the specific effect of the incident disease is not very clear, unless the disease metrics spikes are large enough. For example, in Figure 3, a large spike in cases around weeks 34-36 show a corresponding spike in the search volume activity in the preceding weeks (around week 31-34).

When the time series are examined statistically, they are seen to have stationarity of the residuals, which means that there is a predictable response in the curve for Google search query volume with a change in the number of outbreaks or case counts. However, this is not as strong a relationship as had been previously demonstrated in the studies emanating from developed countries and the rejection of the unit root null hypothesis barely reaches statistical significance for both the outbreak-related model ($p=0.048$) and case count related model ($p=0.041$).

These discrepancies can be explained based on two main issues. The data obtained from the IDSP is probably not representative of the incidence of the diseases in the country. Moreover, because of the method by which the data is collected, this data is primarily sourced from the rural and peri-urban areas of India. However, though mobile connectivity has become ubiquitous in India, Internet penetration has been poor, owing to
lop-sided economic development. Moreover, use of Internet to search for healthcare in-
formation is not a very well-recognized pattern. This behavior is even more restricted in
rural areas, where, owing to lower socioeconomic status, poorer awareness levels and
lower literacy rates, Internet utilization rates remain low.
Considering the limitations of the available data, it would, therefore, be inappropriate
to extrapolate the findings to arrive at concrete conclusions about the utility of the data
for modeling or forecasting or “now-casting” disease outbreaks. The trend, however,
is encouraging because it shows that despite the myriad limitations, the data shows
some consistency, which can be expected to improve once better health statistics are
available and there is penetration of Internet utilization for healthcare information needs
to peri-urban and rural India.

Conclusions
The study shows that the time trend of Google search query volume changes in a pre-
dictable manner with change in the total number of fever-causing infectious disease
outbreaks per week as well as with the total number of cases of fever-causing infec-
tious diseases per week. However, owing to the non-representative nature of the data
on disease incidence and the poor penetration of Internet use for seeking healthcare
information in the rural and peri-urban areas, from which the disease incidence and
outbreak counts are primarily sourced, it is premature to use this model to create a
forecasting or “nowcasting” system to identify infectious disease outbreaks before they
assume epidemic proportions.

Limitations
Penetration of the Internet in all parts of India is poor and use of the Internet for health
information seeking is poorly defined. The disease counts (both cases and outbreaks) are
likely to be an under representation since the Integrated Disease Surveillance Program
(IDSP) is yet to be universalized in India. The disease metrics are based on data which
is predominantly sourced from rural and peri-urban areas, whereas the Internet use for
health information seeking is likely to be higher in the urban areas.

Additional Information

Methods
Google search query volume from India, between January 5, 2014 and De-
cember 14, 2014, was downloaded from the Google Trends Insights website
(http://google.co.in/trends). Google Trends data are scaled and normalized using spe-
cific algorithms to ensure that data remains comparable despite different denominators
(total users using the search engine) across time.
Weekly count of outbreaks and cases of the identified diseases was downloaded from the
publicly available reports on the IDSP website (http://idsp.nic.in). The objective of the
study was to show that there was a relationship between the time trends of the search
query volume and reported disease outbreaks and case counts.
Spearman’s rho was calculated to estimate unadjusted correlation between search ac-
tivities with weekly disease metrics (counts of outbreaks and case counts). Time-series
analysis of search activity and disease metrics was done to ascertain whether they
shared a common stochastic drift. The two-step Engle-Granger method was used to
ascertain cointegration under the assumption that the variables were integrated to the
order of one. In the first step the dependent variable (Google search query volume)
was regressed on a constant and the independent variables (outbreaks of the identi-
fied diseases in model 1 and case counts of the identified diseases in model 2) and the
residuals were calculated. In the next step, the Augmented Dickey-Fuller test was run
on the residuals (without a constant term) under the null hypothesis that the Google
search query volume and disease outbreaks (model 1) and case counts (model 2) were
not cointegrated. Rejection of the null hypothesis would be evidence that the residual is
stationary, that is, the time series are cointegrated, thus proving that the Google search
query trend underwent similar changes with changes in the outbreak numbers (model
1) or case counts (model 2) through time, on a week-by-week basis.
Data was entered into MS Excel and statistical tests for time series data were done using
the open-source statistical software Gretl version 1.9.92.

**Supplementary Material**
Please see https://sciencematters.io/articles/201602000012.

**Funding Statement**
This study did not receive any funding support.

**Ethics Statement**
Not applicable.

**Citations**


