Informatics for Health and Social Care, March 2012; 37(2): 106–124

ISSN 1753-8157 print/ISSN 1753-8165 online
DOI: 10.3109/17538157.2011.647934

informa
healthcare

Syndromic surveillance models using Web data: The case of scarlet fever in the UK

Loukas Samaras[1], Elena García-Barriocanal[2] & Miguel-Angel Sicilia[2]

[1]Ministry of Employment and Social Insurance, General Secretariat of Social Security, Department of National Security Registries and Internet, Athens and [2]Computer Science Department, University of Alcalá, Madrid, Spain

Abstract
Recent research has shown the potential of Web queries as a source for syndromic surveillance, and existing studies show that these queries can be used as a basis for estimation and prediction of the development of a syndromic disease, such as influenza, using log linear (logit) statistical models. Two alternative models are applied to the relationship between cases and Web queries in this paper. We examine the applicability of using statistical methods to relate search engine queries with scarlet fever cases in the UK, taking advantage of tools to acquire the appropriate data from Google, and using an alternative statistical method based on gamma distributions. The results show that using logit models, the Pearson correlation factor between Web queries and the data obtained from the official agencies must be over 0.90, otherwise the prediction of the peak and the spread of the distributions gives significant deviations. In this paper, we describe the gamma distribution model and show that we can obtain better results in all cases using gamma transformations, and especially in those with a smaller correlation factor.

Keywords: *Web data*, *syndromic surveillance*, *scarlet fever*, *gamma distribution*, *Pearson correlation*

1. Introduction

Syndromic surveillance systems are concerned with the continuous monitoring of public health-related information sources and the early detection of adverse disease events. They play an increasingly important role in healthcare systems as concerns for timely response to infectious diseases and epidemic outbreaks are increasing worldwide. More than 75,600 people [23] have died from infectious diseases in the European Union in the last 12 years. Furthermore, the number of deaths from a specific infectious disease such as influenza was 120,176 worldwide [24] during the period 2003–2008, and many thousands of people have suffered more or less serious symptoms. These figures highlight the importance of surveillance systems that integrate as many information sources as possible.

Scarlet fever is a disease caused by infection with the group A *Streptococcus* bacteria [1]. Scarlet fever was once a very serious childhood disease, but is now treatable. There are some known risk factors associated with the infection, including the disease's time of onset (March–June), as described by Wang et al. [2]. We also know that the epidemics of scarlet fever have gone through different phases in the past, as evidenced by the

Correspondence: Miguel-Angel Sicilia, Computer Science Department, Polytechnic Building, University of Alcalá, Ctra. De Barcelona km. 33.6, 28871 Alcalá de Henares (Madrid), Spain. E-mail: msicilia@uah.es

model of scarlet fever in Liverpool (UK) between 1848 and 1900, reported by Duncan et al. [3], based on annual death data. Scarlet fever infection consequently remains an interesting case for surveillance, as outbreaks may arise due to different factors from the past.

Syndromic surveillance relies on the real-time use of information about the population to identify health issues of concern and address them before they become epidemics. As a consequence, a syndromic surveillance system implements a variety of outbreak detection algorithms, requiring a good understanding of the strengths and limitations of various detection techniques and their applicability [4], including data available via the Web and searches of physicians' databases [5].

Traditional syndromic surveillance systems are based on data collected at national or international level by the competent authorities, but a Web surveillance system has not been officially established despite several efforts. At an international level, useful data on various diseases are examined by the World Health Organization (WHO) and European networks such as the European Influenza Surveillance Network (EISN) [16] and the European Influenza Surveillance Scheme (ESSI) [17].

Recent research has investigated the potential of using Web search data to predict the patterns of the spread of infectious diseases. Ginsberg et al. [6] examined the history of Google search queries to 'track influenza-like illness in a population'. That work was based on data from the US Centers for Disease Control and Prevention (CDC) [21] for influenza. The findings of this research resulted in the conclusion that the relative frequency of certain queries is highly correlated with the percentage of visits to a physician in which the patient presents with influenza-like symptoms. Other similar studies have confirmed the hypothesis that Web data represent a reliable and effective source of information for syndromic surveillance. For example, Hulth et al. [7] analyzed the data submitted to a Swedish Web site (www.vardguiden.se) [24] and used partial least squares regression (PLSR) to confirm the reliability of the data. Another study related to an Internet Search Term Surveillance for Flu was reported by Polgreen et al. [8]. They examined the queries submitted to the Yahoo! search engines using linear regression methods. Similarly, Andersson et al. [9] used covariates to interpret the connection of influenza-like illness (ILI) to weekly laboratory diagnoses of influenza (LDI). The work of Johnson et al. [10] is also worthy of mention as a less recent one analyzing Weblogs for similar purposes.

In view of the recent research mentioned above, surveillance systems based on Web data appear to have some potential for complementing other techniques [4] that assist national health plans (e.g. for immunization policies) and International Coordination Systems, such as the European Surveillance Networks (EISN and EISS).

All the studies mentioned above have used various statistical estimation methods and given interesting and promising results. In particular, the study of Ginsberg et al. [6] gives fairly accurate estimations of visits to the doctor using Google data based on log linear models. However, a major problem occurs when the correlation of data is less statistically significant at the level of $\alpha = 0.05$ (95% confidence interval and Pearson correlation factor <0.90), which mainly leads to inaccurate estimations of epidemics' peaks and the spread of a disease, even in a case of a known distribution such as the gamma distribution.

The main focus of this paper is to use data from Web search queries from Google Insights for Search and to highlight the correlation with scarlet fever in the UK using two statistical methods. Therefore, the goals of this research are as follows:

- to use the application and analysis of the statistical model described by Ginsberg et al. [6] for a different disease, that is, scarlet fever, based on data obtained from the Health Protection Agency of the UK and

- to give alternative estimation and prediction models based on gamma distribution; estimation models are used to find correlations between the two kinds of data, and prediction models use estimation models using both a logit model and a gamma distribution model.

Google Trends and Google Insights for Search are useful tools that can provide data from specific Web queries submitted in Goggle search engine. These data were compared with the data from the UK agency in order to determine the correlation between the true development of scarlet fever and the amount of search queries submitted to Google search engines, that is, to find the correlation between what people are searching for and what is really happening.

As a summary of the results, it can be said that if a known distribution can be identified, the estimation is better, especially in cases in which the correlation is not significant enough. When analyzing the case of scarlet fever in the UK for the years 2008, 2009 and 2010, we found that the Pearson R is above 0.90 only in the year 2009. When applying gamma distribution models, taking the highest R (year 2009) as the base, we obtained (1) a better estimation of the peaks of this disease and (2) better correlation factors. The overall figures are given in Table I.

The results of this study highlight the need for testing alternative statistical models for various diseases, as the accuracy of the models is critical in detecting outbreaks.

The rest of this paper is structured as follows. Related work is described in Section 2. The materials used for the study are briefly described in Section 3, with an explanation of their structure and the different methods applied to acquire them. The statistical methods and processes are provided and related to the method described by Ginsberg et al. [6] in Section 4. This section is divided into two parts. In the first part, the analysis of the logit model is provided, and in the second part, the alternative gamma distribution model is presented. The limitations of the two models described are discussed in Section 5. Finally, conclusions and outlook are provided in Section 6, followed by Acknowledgement.

2. Related work

2.1 Johnson et al. [10]

The main purpose of the study by Johnson et al. [10] was to determine whether the level of influenza in a population correlates with the number of times Internet users access information about influenza on health-related Web sites. The data used were gathered from Weblogs, which contain information about the users and the information the users accessed, and are maintained electronically by most Web sites, including HealthLink. These data were produced using weekly counts of the number of accesses of selected influenza-related articles on the HealthLink Web site, and their correlation with traditional influenza surveillance data from the CDC was measured using the cross-correlation function (CCF). Timelines were then defined as the time lag at which the correlation was maximum.

In particular, this study analyzed data from http://www.cdc.gov/ncidod/diseases/flu/weeklychoice.htm during the influenza season (October–May) in the USA, and data sets were obtained from the CDC Web site.

The relevant HealthLink Web articles were identified by reviewing the titles of all the 1,504 articles available on HealthLink in 2001 and 21 articles that users with influenza might consult for information about their illness were identified. The result of this procedure was a set of 17 influenza-related articles that were also assigned into two groups –

Diagnosis/Treatment (11 articles) and Prevention/Vaccination (6 articles). It is also very important to note that a *Perl script* was developed to identify accesses to the 17 influenza-related articles in the Web access logs.

The statistical method used was a *cross-correlation analysis*, using the CCF to measure the correlation between article access counts and influenza activity.

The study came to the conclusion that there was a moderately strong correlation between the frequency of influenza-related article accesses and the CDC's traditional surveillance data, but the results on timelines were inconclusive. Furthermore, using improved methods developed for performing spatial analysis of the data and the continuing increase in Web searching behavior among Americans, the conclusion was that Web article access has the potential to become a useful data source for public health early warning systems.

2.2 Polgreen et al. [8]

The study by Polgreen et al. [8] used search queries from http://search.yahoo.com between March 2004 and May 2008, counting daily unique queries originating in the USA and containing influenza-related search terms. These counts were divided by the total number of searches, and the resulting daily fraction of searches was averaged over the week. The estimation was performed using *linear models*, using searches with 1–10-week lead times as explanatory variables, to predict the percentage of positive influenza cultures and the deaths due to pneumonia and influenza in the USA.

Two types of US influenza surveillance data were used. The first type of data is based on weekly influenza cultures from clinical laboratories throughout the USA which are members of either the WHO Collaborating Laboratories or National Respiratory and Enteric Virus Surveillance System (NREVSS). The second type of data summarizes weekly mortality from pneumonia and influenza. These data were collected from the 122 Cities Mortality Reporting System.

Using the culture data, a linear model was fitted to test the predictability of search frequency on positive influenza cultures, including a time-trend variable (t). The results showed that there was a positive relationship between the fraction of influenza-related queries and positive influenza culture rates two weeks later ($p < 0.001$).

Separate models were also fitted with lags from 1 to 10 weeks for each of the nine US census regions. The results were similar to those of the national model, with the best fitting models predicting positive influenza cultures 1–3 weeks in advance. The average R^2 at 2 weeks was 0.3788. However, values varied from a high value of 0.5729 in the East South Central region to a low value of 0.1656 in the Mid-Atlantic region.

In the second type of data, there was also a positive relationship between the fraction of influenza-related search queries and pneumonia and influenza mortality 5 weeks later ($p < 0.001$). For each of the nine US census regions, the average R^2 at 5 weeks was 0.3041. However, the values varied from a high value of 0.4250 in the East North Central region to a low value of 0.1227 in the Pacific region.

2.3 Andersson et al. [9]

In the study by Andersson et al. [9], the aim was to suggest both simple and advanced rules for how to predict the time and height of the peak of LDI by means of very early observations. The incidences of laboratory-diagnosed cases were compared with the sentinel reporting to examine their relationship and the usefulness of these series for predicting yearly outbreaks.

The data used were on ILIs reported to the Swedish Institute for Infectious Disease Control (SMI) by sentinel physicians, while laboratory-verified infections were reported

by microbiological laboratories. The reporting systems are described at www.smittskyddsinstitutet.se. A Web-based system called Sentinet was developed in Sweden in 2003 for use by both sentinel physicians and laboratories in order to facilitate the submission of data to SMI. The ILI data used in this study covered seven seasons, from the 1999–2000 season to the 2005–2006 season. Data on eight seasons were used for LDI (also including 1998–1999). LDI included influenza cases of both type A and type B.

The incidence of influenza during each season was estimated as a *unimodal regression* on time. The analysis used the methodology of *nonparametric least squares* under the order restriction of unimodality, but without any other assumptions on the regression function. This technique produced consistent estimates for the time and the height of the peak. All analyses were performed by fitting a linear regression to the time of the peak with the estimated early characteristics as independent variables. A similar analysis was carried out for the height of the peak. The predictions were carried out using *covariates* calculated from data in early LDI reports.

The results showed a relation between ILI and LDI that was investigated, and it was found that the ILI variable is not a good proxy for the LDI variable. The advanced prediction rule regarding the time of the peak of LDI had a median error of 0.9 weeks, and the advanced prediction rule for the height of the peak had a median deviation of 28%. Furthermore, through suggested rules for each of the above steps, it was found that this model can give rough predictions as early as about 8 weeks before the peak appears. More advanced prediction rules based on smoothing with unimodal regression giving greater precision have also been suggested.

Finally, the conclusion was that the timing and height of the peak of the yearly influenza season can be reasonably well predicted using available early data and simple rules. Better predictions are achieved using nonparametric regression.

2.4 Ginsberg et al. [6]

The work by Ginsberg et al. [6] aggregated historical logs of online Web search queries submitted to Google search engines between 2003 and 2008 and then computed time series of weekly counts for 50 million of the most common search queries in the USA. Separate aggregate weekly counts were kept for every query in each state for which time series were built. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction.

The model was developed based on data from the CDC's US Influenza Sentinel Provider Surveillance Network. An automated model was designed to examine ILI-related search queries to include the appropriate ones into the model.

This process revealed the top 45 queries to be used in the final model. The number of queries in each topic was indicated, as well as query volume-weighted counts, reflecting the relative frequency of queries in each topic.

Finally, a *linear model* was fitted to weekly ILI percentages between 2003 and 2007 for all nine regions, thus giving a single region-independent coefficient.

The correlation of data was based on *a log linear model* and the findings showed that there is a significant correlation and, in consequence, there is a capability across the nine regions to consistently estimate the current ILI percentage 1–2 weeks ahead of the publication of reports by the CDC's US Influenza Sentinel Provider Surveillance Network.

In conclusion, the study considers that Google Web search logs can provide one of the most timely, far-reaching influenza monitoring systems available today, since the estimates are current every day, while traditional systems require 1–2 weeks for surveillance data to be gathered and processed.

2.5 Hulth et al. [7]

The hypothesis of the work by Hulth et al. [7] was similar to those mentioned above, that is, queries on influenza and ILI would provide a basis for the estimation of the timing of the peak and the intensity of the yearly influenza outbreaks which would be as good as the existing laboratory and sentinel surveillance.

This study was based on the calculation of the occurrence of various queries related to influenza from search logs submitted to a Swedish medical Web site during two influenza seasons (total queries submitted to the search engine were 1,522,802 in 2005/2006 and 2,699,097 in 2006/2007).

The statistical method used was the *PLSR applied in highly correlated data*. The findings of this work showed that certain Web queries on influenza follow the same pattern as that obtained by the two other surveillance systems for influenza epidemics and that they have equal capacity for the estimation of the influenza burden on society. The conclusion was that there is potential for Web queries to be used as an accurate, cheap and labor-extensive source for syndromic surveillance.

By using this statistical approach, which takes various queries into account, and by loading weights of these queries, the results showed that most of the selected types of queries followed the same epidemic pattern as the sentinel and the laboratory influenza data.

2.6 Zhou and Shen [11]

In the study by Zhou and Shen [11], unique search queries submitted to the Baidu search engine in 2008 that contained disease-related search terms were counted. The news articles aggregated by Baidu's robot programs that contained disease-related keywords were also counted. This research found that both the search frequency data and the news count data have a distinct temporal association with disease activity.

The results were based on an examination of the Baidu search database for infectious disease surveillance. With millions of search queries collected in 2008, unique queries associated with infectious-disease-related terms every day were counted and the search frequency data were used as the first data source for predicting disease occurrence. The diseases that were examined were scarlet fever, dysentery, AIDS and tuberculosis, as well as the number of deaths attributable to AIDS and tuberculosis.

To measure disease occurrence, two types of data were used, published by the Chinese CDC (http://www.chinacdc.net.cn/n272562/n276018/index.html accessed on 20 March 2009). The first type of data is based on the number of people infected with scarlet fever, dysentery (including both bacillary dysentery and amebic dysentery), AIDS and tuberculosis, while the second type of data summarizes the monthly mortality attributable to AIDS and tuberculosis.

The statistical method that was used is the *traditional linear regression* analysis method, comparing the correlation between resulting estimates and the disease occurrence measurements.

Since this study examined various diseases, the conclusion included a comparison between them. The conclusion was that the timing and size of epidemic outbreaks vary from disease to disease, hindering efforts to produce reliable and timely surveillance results of different infectious diseases. However, it was found that both the disease-related search frequency and the disease-related news count have a distinct temporal association with disease activity. In addition, the models performed better for some regions than for others, suggesting that events in some regions may increase searches in other regions. The authors of this study also noted that additional work is needed to

examine the spatial relationship between Internet searches and the geographic spread of different infectious diseases.

In general, the model described in this study is considered to be able to publish up-to-date surveillance results, which are about 10–40 days ahead of the release of Chinese CDC reports.

3. Materials

The collection of time series can be extracted from Google Trends [19] and Google Insights. *Google Trends* analyzes a portion of Google Web searches to compute how many searches have been done for the terms entered, relative to the total number of searches done on Google over time. Similarly, *Google Insights for Search* 'analyzes a portion of worldwide Google Web searches from all Google domains to compute how many searches have been done for the terms you've entered, relative to the total number of searches done on Google over time'. They are based on the same data but with different resolutions, and the latter provides additional user tools.

Google Insights was found [18] to be more useful for this study as it provides some additional tools and it enables detection of keywords that are similar to those we are searching for. In particular, Google Insights gives options for year, category and country and useful information on other keywords being searched for and keywords of increasing searches, which may assist in identifying the keywords to search for, especially in large data sets.

Google Insights for Search analyzes a portion of worldwide Google Web searches from all Google domains to compute how many searches have been done for the terms entered, relative to the total number of searches done on Google over time. The results can be seen either by graphs or by downloading a csv file with data, provided that there is a Google account to login to. This downloading can be done from the results page using an Internet browser or by programming a module to connect and download the data, such as a Python script.

We addressed search terms related to 'scarlet fever' in the UK for the years from 2004 to present. We conducted 25 (single or mixed) searches. Table I presents the comparison of correlation factors for every model used.

Table II presents the examined queries from Google Insights, where the column 'keyword' shows the word used in the query. The column 'results' shows the number of different kinds of queries submitted to Google using the keyword concerned. For example, for the keyword 'Scarletina', the number 51 means the 51 different combinations of queries that use this word together with any other word, for example, 'Scarletina Fever', 'Scarletina England', etc. The final column, 'statistical significance' shows whether there is a statistical significance of over 0.90. If so, then we include it in the method. When the value in this column is 'no', this means that the significance is less than 0.90.

We established a requirement that the Pearson correlation factor (R) had to be above 0.90 and the significance level (α) had to be 0.05, so that the results are reliable and

Table I. Comparison of correlation factors (Pearson R).

Year of study	R, using logit model	R, using gamma
2008	0.784281	0.864317
2009	0.901928	0.970211
2010	0.808869	0.890108

Table II. Google Insights search queries used, the results obtained and their statistical significance.

Query number	Keyword	Results	Statistical significance
1	scarlet fever	37	Yes, >0.90
2	scarlet fever + complication	47	No
3	scarlet fever + remedy	81	No
4	scarlet fever + symptoms	49	No
5	terms for + scarlet fever	58	No
6	specific + scarlet fever + symptoms	50	No
7	antibiotic + medication	82	No
8	scarlet fever + antibiotic	37	Yes, <0.90
9	antiviral + medication	78	No
10	scarlet (health)	37	Yes, <0.90
11	scarlet fever (health)	36	Yes, <0.90
12	scarlet	72	No
13	infection + scarlet fever	86	No
14	infectious + disease	78	No
15	scarlet + pharmacy	89	No
16	fever (health)	55	No
17	fever (all)	44	No
18	infectious + fever	46	No
19	therapy	no	No
20	fever + virus	60	No
21	scarlet + fever	52	No
22	scarlet fever rash	39	No
23	scarlet fever + rash	63	No
24	scarlet + fever + rash	59	No
25	Scarletina	51	Yes, <0.90

appropriate for use for estimation and forecasting purposes. We took $R > 0.90$, because as mentioned above, a smaller R does not give a good statistical basis and $a = 0.05$, assuming a confidence interval of 95%, which is widely used in statistical science. The most statistically significant query was 'scarlet fever' and we decided to use only that, as the others did not meet the requirements for reliability.

A Python script was used to gather data from Google Insights. This language can be used in both Windows and Linux-based systems, since it is open source. A script of this kind consists of the following parts:

Part 1. Declarations
This declares the inner Python modules to be used, such as
import cookielib,import os, import urllib, import urllib2, import re, import csv

Part 2. Login parameters
The Google username and password, keywords, the country (geo), etc.

Part 3. Persistent cookie to be saved
"PersistentCookie":"yes",

Part 4. The connection to Google using the login data populate the cookie
login_data = urllib.urlencode(self.login_params).....
and finally

```
Part 5. Obtaining the data and saving it to file as .csv
#to get the data:
self.raw_data=
self.opener.open("http://www.google.com/insights/search/overviewReport?"+params).read
()
# to save the data, e.g. for Spain as spain.csv
fileObj = open("spain.csv","w")
```

The data for scarlet fever were obtained from the UK's National Health Agency (http://www.hpa.org.uk/), with the weekly reports saved in .pdf format. For instance, the 'NOIDS weekly report 0905' [22] is for the year 2009, week 5. Each report includes data for the previous 6 weeks and the notifications of infectious diseases by civilians to doctors, before these are confirmed in the laboratory.

The data from Google Insights include the date for every week (week start–end date) and the proportion of worldwide Google Web searches (in absolute figures) for the specified disease and country. These data were used in this study.

4. Methods and results

Despite the various statistical methods used in previous works, we focused on the model used in the study by Ginsberg et al. [6]. The overall approach was to use a log linear model to confirm a hypothesis. Nevertheless, all the methods were used to confirm an initial hypothesis – in this case the relationship between data from the Web queries and those from the Health Agencies.

All the models in general can be summarized as follows:

- First, a base model is used to retrieve the mathematical relation of data.
- The parameters of the model are then estimated.
- The model is applied to other cases (years, countries and diseases).
- Finally, the results are found and interpreted in order to confirm or reject the initial hypothesis.

4.1. Linear regression model

The general objective of this model is to test the correlation of the query data (Q values) to the probability of scarlet fever (P) using the parameters of the log linear regression.

Here, we state the use of the linear log model $\text{logit}(P) = b_o + b_1 \times \text{logit}(Q) + \varepsilon$, where $\text{logit}(P)$ is equal to

$$\log\left(\frac{P}{1-P}\right) = \log(P) - \log(1-P),$$

which is the logit of the observed data from the Health Agency,

$$\text{logit}(Q) = \log\left(\frac{Q}{1-Q}\right) = \log(Q) - \log(1-Q),$$

which is the logit of the Google query values and b_0 and b_1 are the parameters of the linear model and ε is the error of the estimate.

In statistics, we interpreted the results of this model, as Vasicht [12] did in his paper 'Logit and probit analysis', as follows:

- As P goes from 0 to 1, the logit goes from –00 to +00 and although the probabilities lie between 0 and 1, the logits are not so bounded.
- Although the logit is linear to X, the probabilities are not.
- The interpretation of the model is as follows: b_1 is the slope and b_0 is the intercept.
- If we want to estimate the probability rather than the odds of an event, this can be done directly once the estimates b_0 and b_1 are available.
- Finally, the linear probability model assumes that P is linearly related to every X, in this case to Q, while in the logit model, the log of odds is linearly related to Q.

Applying the above model, we can estimate the parameters of the regression (b_0 = 1.36297415014423 and b_1 = 1.37354944429685), find the logit(P) and finally turn the logit(P) into P using the following equation:

$$P(1 + e^{\log it(P)}) = e^{\log it(P)} \Rightarrow P = \frac{e^{\log it(P)}}{1 + e^{\log it(P)}},$$

which is the estimated P from this model.

The result obtained by applying linear regression is shown in Figure 1.

In Figure 1, p shows the real frequencies of cases, while $p\prime$ shows the estimation made with the logit model using Google query data. The model has a good fit with Pearson R = 0.90192280395531.

Figure 2 shows a residuals chart, in which the residuals are scattered close to the fit line, with no known tendency, and the specific model is, therefore, statistically correct.

Using the above model 1 year backward (2008) and one year forward (2010), we found that these years do not offer a good estimation using this model. The results are apparent in Figure 3.

As shown in Figures 3 and 4, the model describes data well enough only for the year 2009 and not for the other 2 years. It can be seen that the peak of the real data differs a great deal from the estimated values. The reason for this is the lower Pearson R factors, which are even less than 0.80, as described in the introduction.

4.2. The gamma distribution model based on differences

After considering the above estimates, we understood the difficulties, but still had one important finding: the data for every year seemed to follow a gamma-like distribution. If

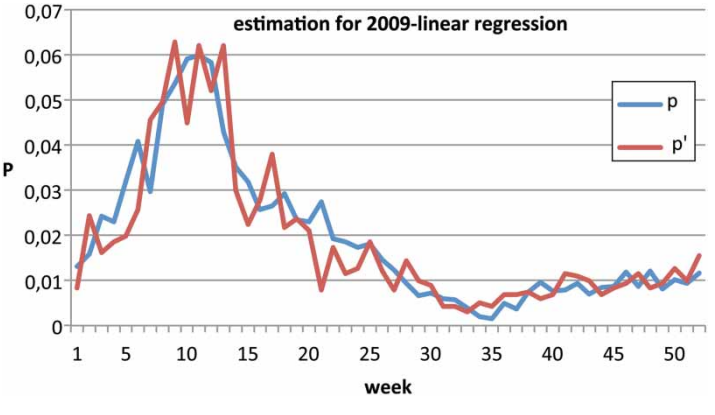

Figure 1. Estimation of the log linear model (logit) for the year 2009.

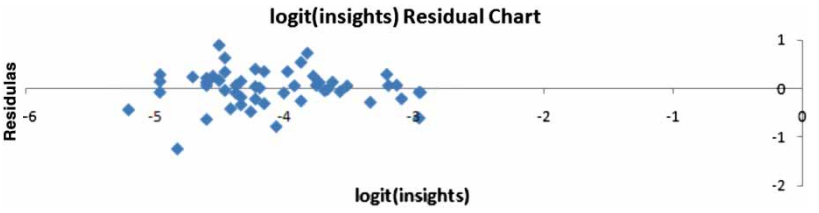

Figure 2. Residuals of the model.

we could ascertain it for the year 2009, the gamma transformation gives a good estimate, and we can use this distribution as the basis for evaluating the other years.

The general objective of this model is to estimate the probability of scarlet fever of a gamma distribution (f) from the search query data (x), using the parameters of the gamma distribution.

A general gamma distribution (the probability density function, f) can be found using the following equation:

$$f(x;k;\theta) = x^{k-1}\frac{\mathrm{e}^{-x/\theta 2}}{\theta^{\kappa}\Gamma(\kappa)}, \quad x \geq 0, k, \theta \succ 0,$$

where

$$\Gamma(\kappa) = (\kappa - 1)! \quad \text{for the positive integer of } k,$$

or generally $\Gamma(k) = \int_0^{\infty} t^{s-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$

Since we were modeling time series, we used the cumulative density function (F) and specifically the three-parameter gamma (Gamma-3P CDF), applying it to the following equation:

$$P' = [G(\mathrm{Google}_y) \cdot G(\mathrm{df}_{2009}) \cdot \mathrm{Adj}_y] + c,$$

where $P\prime$ is the predicted probability of the model for scarlet fever, $G(\mathrm{Google}_y)$ are the values of Google data corresponding to the gamma distribution of the year y and $G(\mathrm{df}_{2009})$ are the values corresponding to the gamma distribution of the differences between the observed and expected values for the year 2009, which is the year with the

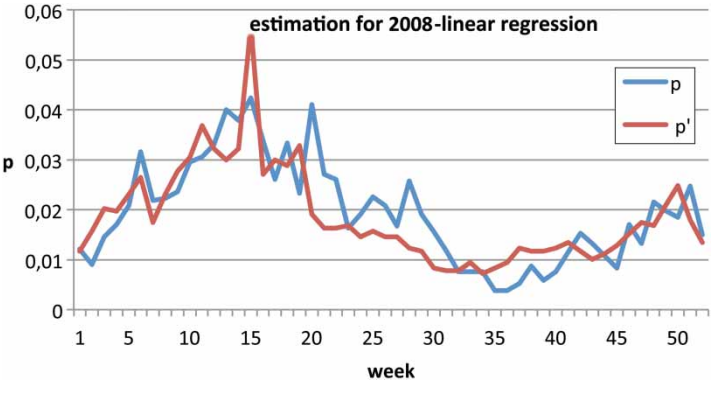

Figure 3. Estimation of the log linear model (logit) for the year 2008.

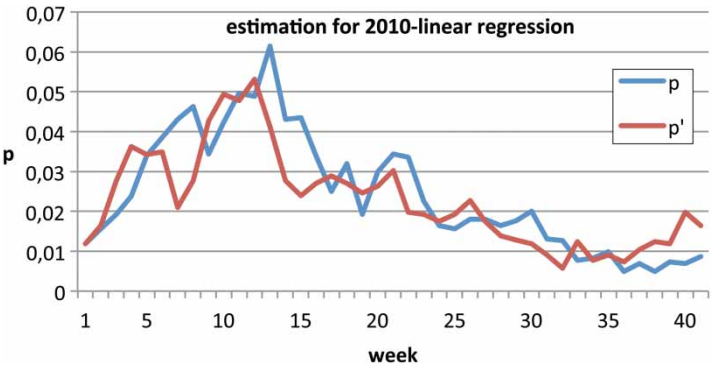

Figure 4. Estimation of the log linear model (logit) for the year 2010.

highest Pearson factor.

Adj_y is the adjustment factor for every year y. It is given as follows:

$$\text{Adj}_y = 1 - \left[\frac{\ln(b_y) - \ln(a_{2009})}{8}\right],$$

where $\ln(b_y)$ is the natural logarithm of the b-parameter of gamma distribution of the Google values for each year y and $\ln(a_{2009})$ is the natural logarithm of the a-parameter of gamma distribution of Google values for the year 2009. The scale adjustment values must, therefore, be calculated as follows:

$$c = \frac{1 - \sum_{w=1}^{26} p}{n - 6}, \quad n \notin \{\max(n), \max(n-5)\},$$

where c is the scale adjustment value, since the sum of p-probabilities is less than 1 for years other than 2009 and for the all the n-weeks, except for the week with the peak and the 5 weeks before it.

The above model assumes that the differences for every year are equal to the X adjustment coefficient based on the observed data from Google plus c.

Applying this model for the year 2009, we obtained the results shown in Figure 5.

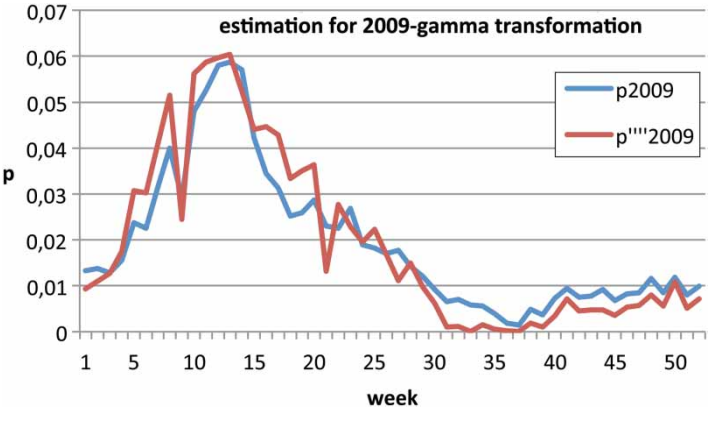

Figure 5. Estimation for the year 2010 using gamma.

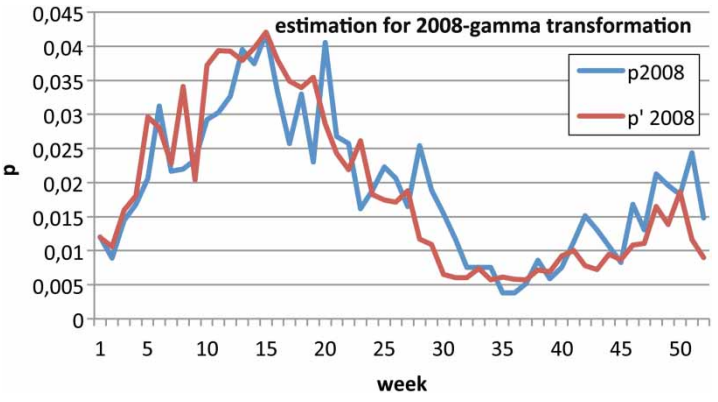

Figure 6. Estimation for the year 2008 using gamma.

It is obvious that the adjustment coefficient of the year 2009 is equal to 1 and $c = 0$.

From this figure, we can make a very good estimation both for the peak of the disease and for its spread. In statistical terms, we obtained a better Pearson correlation factor (than the logit model) of 0.970211. This means that we can use this estimation as the basis for other years.

Applying this model for the other years, while knowing nothing about the distribution of the cases, but only the data from Google, we obtained the results shown in Figure 6, and for the year 2010, we obtained the results shown in Figure 7.

It is obvious that the estimation for the years 2008 and 2010 is better when this model is used, since we obtained better probability peaks and better Pearson factors, as mentioned above.

To conclude this analysis, we present the table of distributions in Table III.

The following should be considered at this point:

1. The identification of a known distribution cannot always be successful, but when it exists, it contributes to simplifying the procedure and possibly to a better Pearson *R*.
2. The adjustment coefficient is 0.70 for the year 2008, and it is 0.87 for the year 2010. This means that for every year with the peak of probability being less than or equal to the peak of the year 2009, this factor should be expected to be between 0.5 and 1, yielding

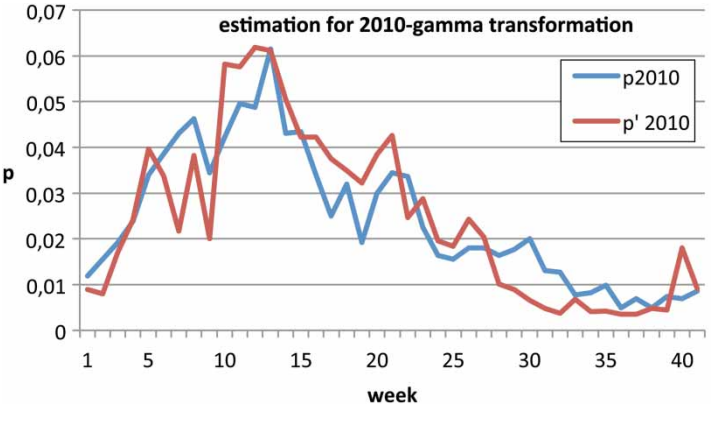

Figure 7. Estimation for the year 2010 using gamma.

Table III. Table of distributions.

	Data								
	2009			2008			2010		
Parameters	*a*	*b*	γ	*a*	*b*	γ	*a*	*b*	γ
Scarlet fever	1.3117	56.426	5.254	3.3782	16.61	−0.1313	0.83389	45.913	12
Time	30.585	2.745	−57.512	30.585	2.745	−57.512	30.585	2.745	−57.512
Google	1.4665	19.392	10.562	1.3974	14.562	12.689	4.0567	10.233	13.071
df	1.6784	33.262	−15.557	4.082	9.6031	−16.258	1.8681	0.00855	−0.00424
Scarlet fever/df	1.279561	0.58948	−2.96098	1.208336	0.578152	123.8233	2.240224	0.000186	−0.00035

lower values as the Pearson R falls. For the possibility of a higher peak than that of the year 2009, the factor is expected to be greater than 1.

3. Assuming a distribution or not results in a procedure of prediction, where the predicted value for every P_t, $t > 0$, depends on previous estimates, no matter how these are made. This means that every subsequent step of prediction depends on a previous series of values, that is, a bound probability.

4.2.1 Prediction

The above hypothesis has the following advantages and capabilities:

- It predicts the spread of scarlet fever on the basis of a known distribution.
- It predicts future values for the weeks after the 41st week.

Nevertheless, it is an a posteriori model based on the knowledge of the peak. In terms of prediction, what happens if we want to perform a prediction before the peak arrives? In this case, we could establish prediction models 5 weeks earlier before the peak is known, that is, in the 8th week. We, therefore, do not know beforehand the parameters of the distribution and the peak. Instead, we only know the data for the first few weeks.

In order to achieve a reliable prediction model, we have to complete the following tasks:

1. estimate the peak and
2. for the weeks after the 8th week.

To estimate the peak for the years 2008 and 2010, a nonparametric model should be used as follows:

$$\max(y) = P_{(\text{google},8)} \frac{\sum_{w=4}^{8} P_{\text{scarlet},w} / P_{\text{google},w}}{4} \cdot \sqrt{3},$$

where max(y) is the peak for the year y and $P_{(\text{google},8)}$ is the google value of the 8th week.

To construct the prediction model, we used the following:

$$P' = [\max \cdot G(\text{Google}_y) \cdot G(\text{df}_y) \cdot \text{Adj}_y] + c,$$

where $P\prime$ is the predicted number of cases and $G(\text{Google}_y)$ are the values from the gamma distribution of the Google values for the first 8 weeks for the weeks after the 8th week the corresponding values for the year 2009 are used (as basis). $G(\text{df}_y)$ are the values from the gamma distribution of the differences (Cases – Google values) for the first 8 weeks. For the weeks >8, the corresponding values for the year 2009 are used (as basis). Adj_y is the adjustment factor of the distribution. Its value is always 2 and is applied only in the 9th week. c is the smoothing constant, which is as follows:

Week	c
9–21	1
22–47	$\frac{P_{(\text{scarlet},w=52)}}{2} = 20.54$
47–52	$P_{(\text{scarlet},w=52)} = 41$

where P(scarlet, $w = 52$) is the value of scarlet fever for the last week of the year 2009.

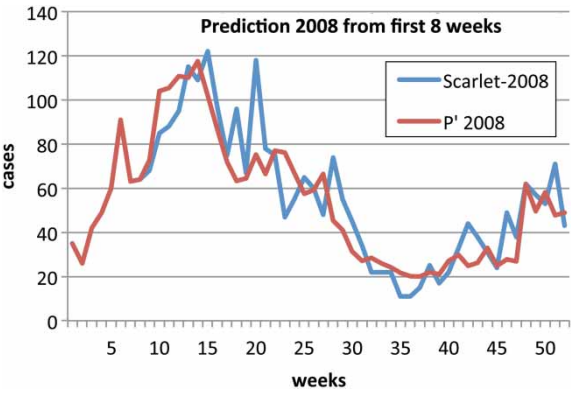

Figure 8. Prediction for the year 2008.

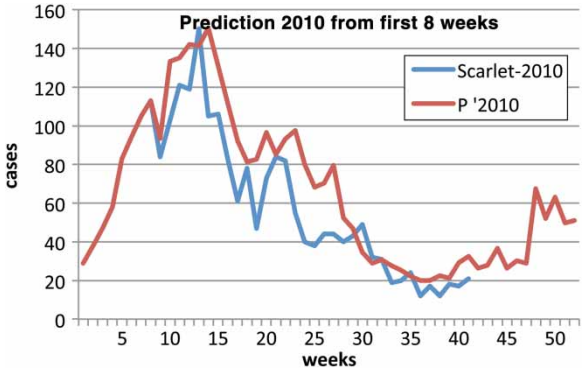

Figure 9. Prediction for the year 2010.

By building this model, and only having data for the first 8 weeks, we can make a prediction for both the peak and the spread and for every week after the 8th week, before we even obtain the future data. Since the peak occurs in the 13th (or 15th) week, our prediction model gives an estimation 5–7 weeks earlier.

The results are shown in Figures 8 and 9.

The correlation factors are given in Table IV.

5. Limitations

At this point, we must refer to the limitations of the models used:

a As regards the log linear model, the best goodness of fit can be obtained on data with high correlation factors ($R > 0.90$), as mentioned above.

b There are two requirements for the gamma distribution model:
- the identification of a gamma distribution and
- the existence of at least one distribution with $R > 0.90$ to be used as a basis,

c The limitations of using data from Web queries are related to the technology's public penetration and the nature of the disease being studied. In specific terms, the limitations are as follows:

Table IV. Correlation factors and predicted peaks.

Year	Pearson *R*	Predicted peak	Real peak
2008	0.891856	115,735	122
2010	0.92926	150,658	150

- The size of the data. The bigger the data set, the greater the number of different estimations that can be made; for example, for influenza it is best to make estimations across regions of relatively big countries, such as the USA. However, for scarlet fever, a smaller data set cannot be used in the analysis.
- For influenza, we can obtain large data sets for many countries, but public surveillance of other diseases such as scarlet fever provides limited data for limited periods.
- The Internet penetration among the public must be high enough for statistically significant data to be retrieved.
- The differences in symptoms in some cases are significant; for example, an online search for a serious situation would be of less importance, given that a serious illness must be dealt with directly by a doctor, probably at a hospital. However, a less serious disease could have a great deal of references on the Web.

6. Conclusions and outlook

Web query data have an interesting potential as a source of syndromic surveillance. In this paper, we have provided models for scarlet fever that provide evidence on this potential, complementing previous research.

In this study, and by means of a pattern that follows a known distribution, using the combination of parametric and nonparametric methods, we predicted the peak and the spread of scarlet fever, *5 weeks* before the arrival of the peak. This model can also be applied, even in cases with a lower correlation factor (Pearson *R*), and it can provide predictions for the spread of the disease for the whole time series of each year.

As a consequence, a Web surveillance system could provide immediate information on the spread of a specific disease, not only within a single country, but also between many countries, before these countries certify their data. In the European Union, progress has been made on the influenza surveillance system over the years, but not for all the other diseases in a similar way. Indicators [14] for successful monitoring have been created, but measurement is different across countries internationally; for example, for ILI visits of possible influenza patients, the US competent authority (US CDC) uses the proportion of outpatient visits for ILI as a percentage, while in Europe, the scale is measured using ILI per 100,000 cases (EISN) and the practice in the past was to use ILI per 1,000 people (ESSI), a system still used in Greece [20] alongside the new European procedure. Furthermore, a relative guide by the European Center for Disease and Control (ECDC) [15] includes a section on planning, preparation and practice for anti-viral assistance between local services within 48 h.

In conclusion, in the subsidiary indicators proposed under the section Key Indicator 1, there is a reference to data based on clinical samples (1.2.a) and structured sampling (1.2. b.). Although additional networks are not included, what does the sample analysis mean? It means that the statistical process for immediate action purposes requires immediate estimation. In the language of statistics, this stands for sample estimation and analysis. We believe that data from both sample data sets from the official agencies and also from

the Internet (e.g. Google Insights) can provide such a quick and parametric or nonparametric estimation.

Acknowledgement

We would like to thank Mrs Sarah Collins at the Health Agency of UK for helping to obtaining the necessary data on scarlet fever.

Declaration of Interest: The authors report no conflict of interest. The authors alone are responsible for the content and writing of the paper.

References

1. Bisno AL, Stevens DL. Streptococcus pyogenes. In: Mandell GL, Bennett JE, Dolin R, eds. Principles and practice of infectious diseases. 7th ed. Philadelphia, PA: Elsevier Churchill Livingstone; chap. 198; 2009.
2. Wang J, Zhang J-Q, Pan H-F, Zhu Y, He Q. Epidemiological investigation of scarlet fever in Hefei City, China, from 2004 to 2008. Tropical Doctor 2010;40:225–226.
3. Duncan SR, Scott S, Duncan CJ. Modelling the dynamics of scarlet fever epidemics in the 19th century. European Journal of Epidemiology 2000;16(7):619–626.
4. Ping Y, Hsinchun C, Zeng D. Syndromic surveillance systems. Annual Review of Information Science and Technology 2008;42(1):425–495.
5. Jormanainen V, Jousimaa J, Kunnamo I, Ruutu P. Physicians' database searches as a tool for early detection of epidemics. Emerging Infectious Diseases 2001;7(3):474–476.
6. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012–1014.
7. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLoS ONE 2009;4(2): e4378:1–16.
8. Polgreen P, Chen Y, Pennock D, Nelson F. Using Internet searches for influenza surveillance (Internet search term surveillance for flu). Clinical Infectious Diseases 2008;47:1443–1448.
9. Andersson E, Kühlmann-Berenzon S, Linde A, Schiöler L, Rubinova S, Frisén M. Predictions by early indicators of the time and height of the peaks of early influenza outbreaks in Sweden. Scandinavian Journal of Public Health 2008;36(5):475–482.
10. Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, Dowling J, Barnas G, et al. Analysis of web access logs for surveillance of influenza. Medinfo 2004;11:1202–1206. Available from: http://books.google.com/books?hl=en&lr=&id=S6Qj6n7rhowC&oi=fnd&pg=PA375&dq=Johnson+%22Analysis+of+web+access+logs+for+surveillance+of+influenza.%22&ots=Q9dfeP00a9&sig=mmezD0jjFedJh3Q9rgO_9z2uhMg#v=onepage&q=Johnson%20%22Analysis%20of%20web%20access%20logs%20for%20surveillance%20of%20influenza.%22&f=false.
11. Zhou X, Shen H. Notifiable infectious disease surveillance with data collected by search engine. Journal of Zhejiang University – Science C 2010;11(4):241–248.
12. Vasicht AK. Logit and probit analysis. Indian Agricultural Statistics and Research Institute, ebooks, Advances in Data Analytical Techniques: E-Book, Module VI-6.5. Available from: http://www.iasri.res.in/ebook/EBADAT/6-Other%20Useful%20Techniques/5-Logit%20and%20Probit%20Analysis%20Lecture.pdf (accessed 29 December 2011).
13. Jewell NP. Statistics for epidemiology. Chapman & Hall/CRC; 2004, CRC Press LLC, 2000 N.W.Corporate Blvd, Boca Raton, Florida 33431.
14. ECDC Pandemic Preparedness Self Assessment Indicators – Version Summer 2007.
15. European Center for Disease Prevention & Control, EU (2009). Technical report: guide to public health measures to reduce the impact of influenza pandemics in Europe: the ECDC menu. Available from: http://www.ecdc.europa.eu/en/publications/Publications/0906_TER_Public_Health_Measures_for_Influenza_Pandemics.pdf (accessed 29 December 2011)
16. European Influenza Surveillance Network (EISN). EISN – weekly electronic bulletins. Available from: http://www.ecdc.europa.eu/en/activities/surveillance/EISN.
17. European Influenza Surveillance Scheme (ESSI). Euroflu weekly electronic bulletins. Available from: http://www.eiss.org/, ECDC Pandemic Preparedness Self Assessment Indicators - Version Summer 2007.

Available from: http://www.ecdc.europa.eu/en/healthtopics/Documents/0705_Pandemic_Influenza_Preparedness_Indicators.pdf
18. Google Insights for Search. Overview. Available from: http://www.google.com/support/insights/bin/topic.py?hl=en&topic=13973.
19. Google Trends. About Google Trends. Available from: http://www.google.com/intl/en/trends/about.html.
20. Greece – Center for Disease Control and Prevention – KEELPNO. Influenza weekly reports. Available from: http://www.keel.org.gr/.
21. United States Center for Disease Control and Prevention – CDC (2004–2010). Weekly reports. Available from: http://www.cdc.gov/.
22. UK Health Protection Agency (2008–2010). NOIDS weekly reports. Available from: http://www.hpa.org.uk/hpr/.
23. World Health Organization Regional Office for Europe. European Detailed Mortality Database (2010). ICD-10 (International Statistical Classification of Diseases and Related Health Problems, 10th revision). Available from: http://data.euro.who.int/dmdb/.
24. World Health Organization (2010, July). WHO Statistical Information System (WHOSIS), WHO Mortality Database. Available from: http://www.who.int/whosis/mort/download/en/index.html.