Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks – results for 2009-10

A Valdivia (tonyvald@hotmail.com)[1], J López-Alcalde[2], M Vicente[3], M Pichiule[4], M Ruiz[5], M Ordobas[6]
1. Preventive Medicine Unit, Hospital de Dénia (Marina Salud), Dénia, Spain
2. Health Technology Asessment Unit, Agencia Laín Entralgo, Madrid, Spain
3. Primary Health Care, Area 8, Madrid, Spain
4. Preventive Medicine Service, Hospital Universitario de La Princesa, Madrid, Spain
5. Primary Health Care, Area 11, Madrid, Spain
6. Epidemiology Service, General Subdirection for Health Promotion and Prevention, Madrid, Spain

Citation style for this article:
Citation style for this article: Valdivia A, López-Alcalde J, Vicente M, Pichiule M, Ruiz M, Ordobas M. Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks – results for 2009-10. Euro Surveill. 2010;15(29):pii=19621. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19621

Article published on 22 July 2010

The number of Internet searches has recently been used by Google to estimate the influenza incidence in the United States. We examined the correlation between the *Google Flu Trends* tool and sentinel networks estimates in several European countries during the 2009 influenza A(H1N1) pandemic and found a good correlation between estimates and peak incidence timing, with the highest peaks in countries where Internet is most frequently used for health-related searching. Although somehow limited, Google could be a valuable tool for syndromic surveillance.

Introduction

On 21 April 2009, the Centers for Disease Control and Prevention (CDC) alerted the media regarding the isolation of the 2009 pandemic influenza A(H1N1) virus from humans. The World Health Organization (WHO) made the unprecedented decision to announce a level 4 pandemic alert on 27 April, raising it to level 6 on 11 June given the strong and sustained transmission of the virus around the world [1].

In the northern hemisphere, surveillance of the pandemic was maintained throughout 2009 via the exceptional use of sentinel physician networks (SPNs) during the summer season. The majority of the European countries reported the weekly incidence of influenza-like illness (ILI) or acute respiratory infection (ARI) through this system [2]. Although such networks allow the rapid and precise collection of information, the average delay between receiving it and its dissemination via epidemiological surveillance websites is about two weeks [3]. In addition, for a case to be registered, contact has first to be made with the health system. These problems have led to investigations into the use of alternative surveillance systems capable of registering more cases in the earlier stages of epidemics, such as recording the number of absentees from work or school, the demand for medications, or the use of Internet surveys [3].

The number of Internet searches made using Google (http://www.google.com) employing search terms related to influenza has recently been used to construct a model for the estimation of influenza incidence in the United States (US). The estimates this model provides correlate very well with SPN data, and can be made available one or two weeks earlier than CDC surveillance reports [4], although the correlation of the model with positive influenza tests is somehow weaker [5]. Currently, estimates are available for 20 countries, 14 of which are European, and can be referred to via *Google Flu Trends* (GFT) at http://www.google.org/flutrends [6].

For Australia and New Zealand, a good correlation has been recorded between the incidence estimates of this GFT model and the sentinel physician networks (SPN) data during the 2009–10 influenza season [7,8]. This period falls between influenza seasons in the northern hemisphere, a time during which discrepancies have been noted in GFT and SPN incidence estimates for the US [9]. In this report we aim to examine the correlation between GFT and SPN incidence estimates in different European countries during the 2009 influenza A(H1N1) pandemic, i.e. both before and during the influenza season. The association between online search habits in each country and the correlations observed were also investigated.

Materials and methods

The weekly (23 March 2009–28 March 2010) GFT and SPN (based on ILI or ARI data) estimates of influenza incidence were recorded for 13 European countries. The sources of the SPN information were the European Influenza Surveillance Network of the European Centre for Disease Prevention and Control (ECDC) [10], the World Health Organization [2], the Réseau Sentinelles de France [11], the Spanish Red Nacional de Vigilancia Epidemiológica [12], Robert Koch Institute (Germany)

[13], and Smittskyddsinstitutet (Sweden) [14]. Spearman correlation coefficients between the GFT and SPN estimates were calculated for the periods before and after 31 August 2009 (i.e. before and during the influenza season) for each country. The influence of the percentage of the different populations making health-related Internet searches (obtained from Eurostat) [15] on the strength of the correlation between the GFT and SPN results was also examined by Spearman analysis. Significance was set at $p<0.05$. All calculations were made using Stata 9.1 software.

Results

The Table shows the correlations between the GFT and SPN (ILI or ARI) results for each country and period examined.

Austria was not included in this analysis because the available data were insufficient. In most countries the correlation was stronger during the second period (i.e. after 31 August 2009), the exceptions being Russia and Ukraine. The two systems commonly coincided in terms of registering peak incidence, although the GFT data sometimes identified this to occur one or two weeks earlier, e.g. for Poland and Switzerland. The two notable exceptions to this were Sweden, for which the GFT model estimated peak incidence to have occurred some 11 weeks before that suggested by the SPN system, and Bulgaria, for which the SPN system suggested a peak incidence one week before the GFT estimate.

Figures 1 and 2 show the SPN ILI and ARI results separately in comparison with the corresponding GFT results. In the majority of cases, the graphs are similar.

The graphs compare the weekly proportion of consultations for acute respiratory illness according to sentinel physician networks and incidence estimates obtained from *Google Flu Trends*. The first week of the series was 23–29 March 2009 (epidemiological week 13).

However, the height of the incidence peaks for France and Hungary appears to be overestimated by the GFT model, and underestimated for Switzerland and Spain (preceded by an overestimation during the summer months in Spain).
Figure 3 shows that the greater the proportion of the population that sought health information via the Internet in 2009, the better the correlation between the GFT and SPN ILI results (Rho=0.7545; p=0.0305). This association was maintained after adding the information from countries that record only ARI data (Germany and Bulgaria) (Rho=0.6991; p=0.0245). The graph shows the correlation between the proportion of individuals who used the Internet for seeking health information in 2009 and the Rho coefficient between the SPN ILI per 100,000 population and GFT incidence estimates.

Discussion and conclusions

In general, the GFT and SPN results (both ILI and ARI) showed a strong correlation. This is the first study to relate GFT and SPN estimates in Europe; the only other northern hemisphere study was undertaken by Doornik

Table

Correlation between weekly sentinel physician network data on influenza-like illness or acute respiratory illness and *Google Flu Trends* incidence estimates

COUNTRY	SYNDROME	CORRELATION			
		Overall period[a]	Pre-epidemic[b]	Epidemic[c]	Peak incidence
		Spearman Rho	Spearman Rho	Spearman Rho	(GFT versus SPN)
Belgium	ILI	0.7358	0.6929	0.8533	Same week
France	ILI	0.9124	0.4957	0.9678	Same week
Hungary	ILI	0.8959	0.3931	0.7496	Same week
Netherlands	ILI	0.8597	0.7850	0.9384	Same week
Norway	ILI	0.8769	0.8651	0.8606	Same week
Poland	ILI	0.7157	0.5179	0.5840	1 week before
Spain	ILI	0.7331	0.6443	0.9471	Same week
Sweden	ILI	0.7733	0.5451	0.8704	11 weeks before
Switzerland	ILI	0.8501	0.7800	0.8783	2 weeks before
Bulgaria	ARI	0.8377	0.6263	0.7260	1 week after
Germany	ARI	0.9396	0.7370	0.9029	1 week before
Russian Federation	ARI	0.8479	0.8149	0.6899	1 week before
Ukraine	ARI	0.8144	0.7875	0.5275	Same week

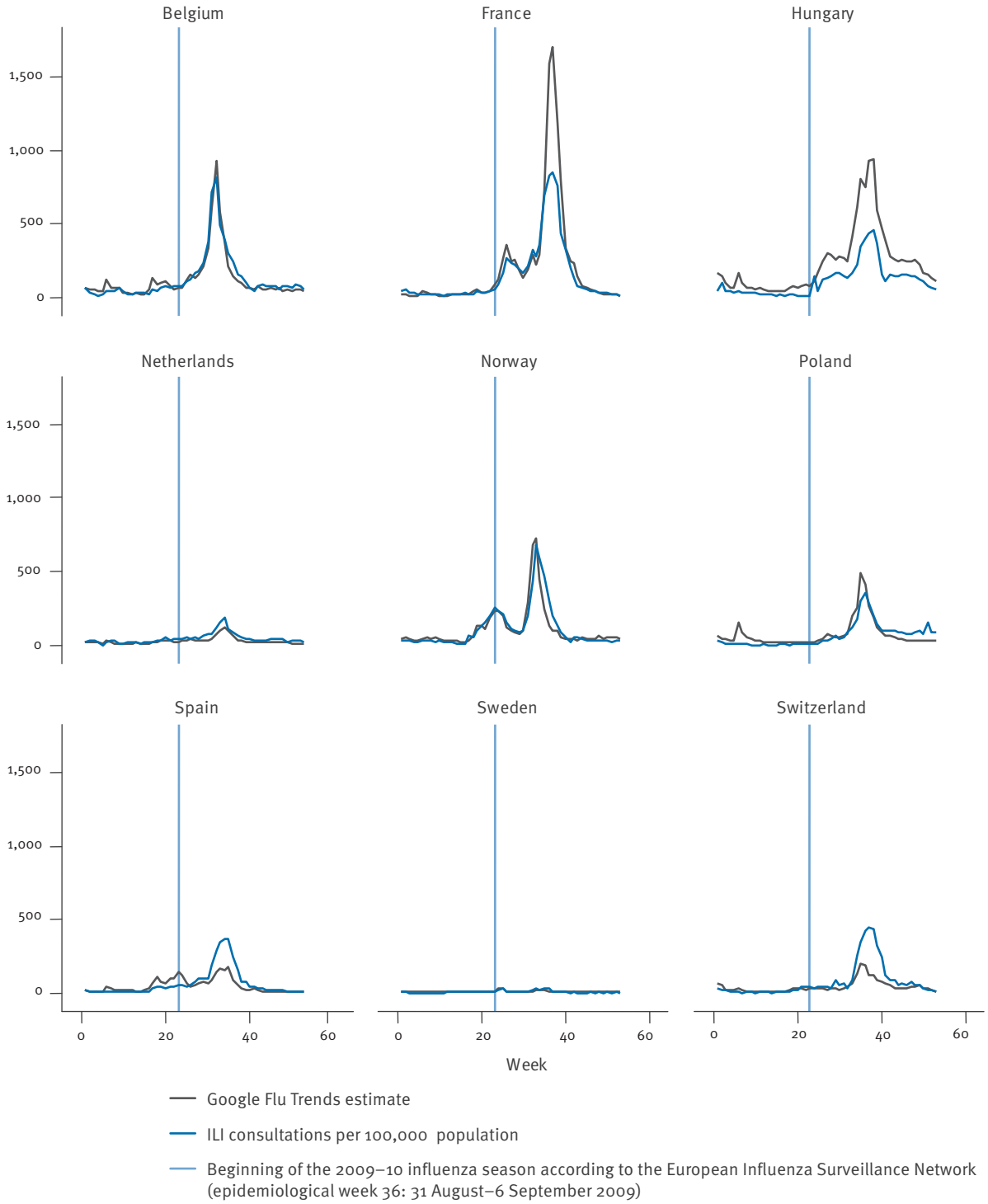
[a] 53 epidemiological weeks: 23 March 2009–28 March 2010.
[b] 23 epidemiological weeks: 23 March 2009–30 August 2009.
[c] 30 epidemiological weeks: 31 August 2009–28 March 2010.
GFT: Google Flu Trends.
SPN: Sentinel Physician Network.

in the US [9], with which the present results are in general agreement. To our knowledge, data from search queries in Google have also been correlated with SPN estimates for chickenpox [16,17] and gastroenteritis [16], showing a similar or higher correlation than ILI.

We made a division into pre-influenza season and influenza season because in the pre-influenza season Internet interest in influenza is likely to be driven mostly by the global interest in a possible pandemic, which may be unusually high and not related with a real increase in the incidence rate of influenza. According to this hypothesis, the correlation observed in the present work was weaker in the period before 31 August than after this date. This might also be related to a lack of incidence data for the summer. The

Figure 1

Weekly influenza-like illness consultations per 100,000 population compared to *Google Flu Trends* estimates of influenza incidence in nine European countries, 23 March 2009–28 March 2010

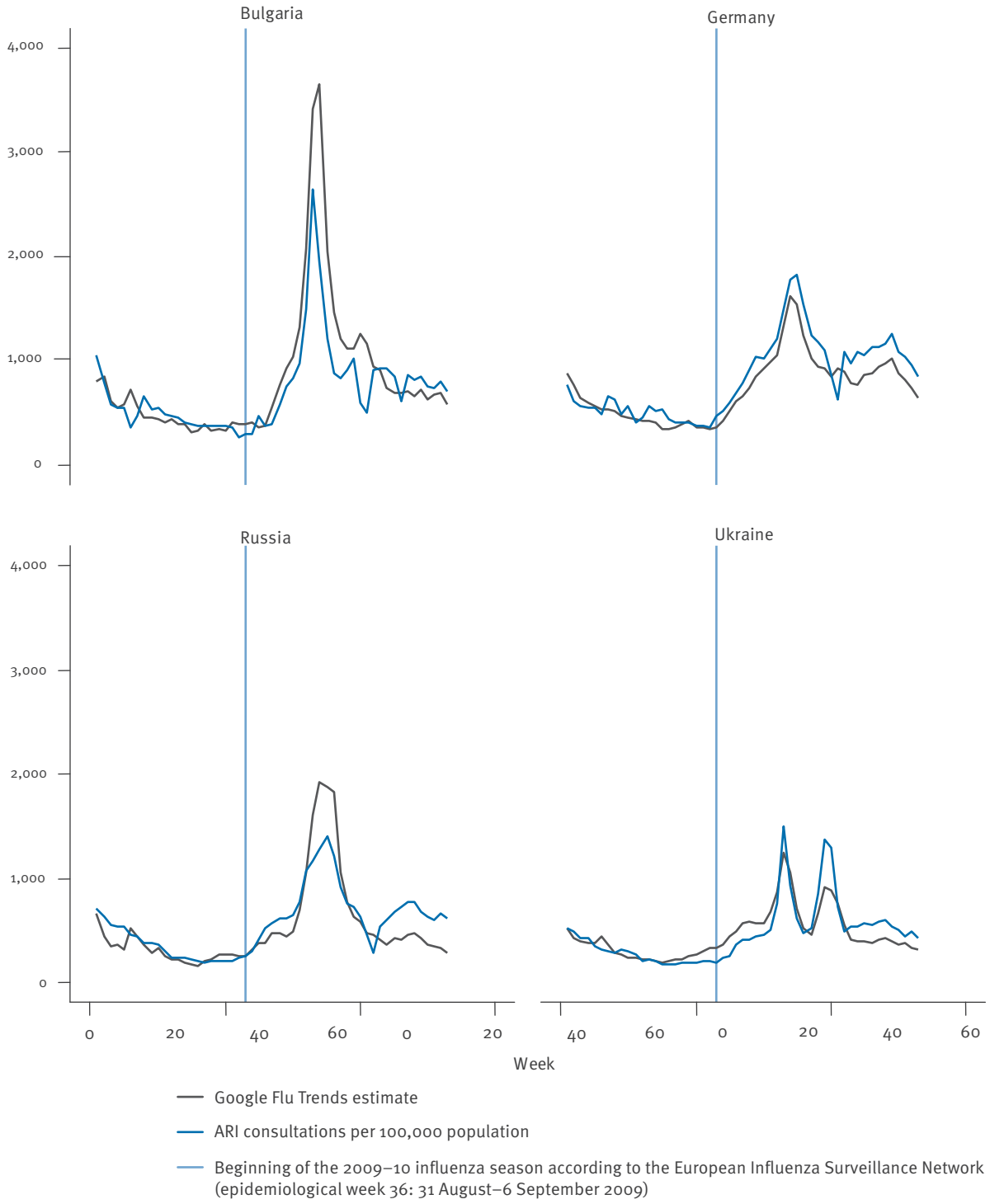


ILI: influenza-like illness.

GFT model is known to provide more robust estimates when incidence rates are higher [4,6]. Nonetheless, in agreement with that indicated by Ginsberg *et al.* [4], the present GFT incidence results were not unduly affected by large numbers of searches for information made before 31 August, i.e. when true influenza incidence was low, probably for the method used by GFT [4]. Google engineers designed an algorithm that detects the search terms most related with ILI, testing the regional variation of Google queries against the regional variations in SPN ILI data. The search fractions for these queries are pooled together in a single search fraction for each week that is used to fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query. The number of top-scoring queries to be pooled together is optimised at estimating out-of-sample points during cross-validation [4]. The Internet Protocol address is used to

Figure 2

Weekly acute respiratory illness consultations per 100,000 population compared to *Google Flu Trends* estimates of influenza incidence in four European countries, 23 March 2009–28 March 2010



ARI: acute respiratory illness.

identify the countries that generate the queries, thus allowing the application of the general method to generate estimates for each single country. This method avoids overfitting using a single explanatory variable and makes the model resilient to variations in only few terms. For instance, at the beginning of the pandemic, there was a massive peak in the search fraction for the term 'influenza' translated in the official languages of each country. This was observed throughout Europe, nonetheless the GFT estimates did not change and continued to be related with the SPN estimates. The only exceptions to this were seen in Belgium, Hungary and Poland (Figure 1).

We used a non-parametric test for the statistical comparison between GFT and SPN estimates. This approach loses information and largely ignores time, but was preferred due to the distribution of the GFT and SPN estimates, significantly different from normal for almost all countries (skewness and kurtosis test, $p<0.05$). In addition, the period considered was too short to justify a multivariate time series approach (e.g. Poisson or binomial negative regression). Thus, we preferred a mixed statistical and graphical approach.

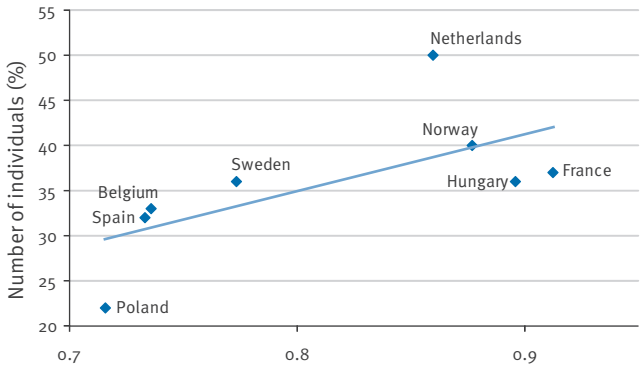
Although the GFT and SPN disease incidence peaks generally coincided or differed by 1-2 weeks (GFT providing an earlier peak in such cases), the GFT peak estimate for Sweden preceded the ILI peak by 11 weeks. This could be related with the sentinel network scheme of Sweden, that presents a lower probability of symptomatic patients to contact a sentinel physician, making ILI estimates less valuable than those from other countries. Large differences were seen in the height of the peaks recorded by each system in France, Switzerland, Hungary and Spain. In addition, in Spain, discrepancies in terms of incidence magnitude appeared during the summer months. This was also reported in the US study, for which correction was made using an autoregression method [9]. This allowed much more robust estimates to be made without losing the capacity to release information one or two weeks before the official CDC reports [9]. The same type of correction might be useful when dealing with European data, in which discrepancies might be the result of different national pandemic control policies or the characteristics of national health and SPN systems. This timely information could be valuable to allocate resources in advance of an epidemic peak, allowing an effective response to sudden changes in the incidence of influenza.

When describing the GFT model, Ginsberg *et al.* [4] indicated that it might be used with good results in any country with a large population of Internet users whose members make regular web searches. The association observed in the present work between the proportion of the population making health-based Internet searches and the strength of the GFT/SPN correlation is in line with the results according to which the strongest GFT/SPN correlations were found in countries where the Internet is more often used as a source for seeking health information. The selected indicator of Internet use in each country (proportion of population that sought health information via the Internet in 2009) describes the health-oriented search habits better than other indirect indicators frequently used (e.g. proportion of households with Internet access, or Internet use at work). The sample (general population of each country) and the period selected (yearly data about Internet use) are representative of the behaviour of European population in the year 2009, and probably highly correlated with the influenza-related searching behaviour during the pandemic.

In conclusion, when disease incidence was high, estimates of the latter based on the GFT model were very similar to those based on SPN data. The GFT model appears robust and could help in epidemiological surveillance by providing more rapid estimates of incidence, i.e. before publication is possible using conventional methods. GFT estimates could well improve in the coming years as actual observations are used to fine-tune the model, and as the use of Internet for finding health information increases. Although the GFT model cannot replace conventional surveillance methods like virological surveillance schemes [5], it may certainly be able to complement them.

Figure 3

Individuals who searched the Internet for health-related information plotted against the correlation between the sentinel physician network/*Google Flu Trends* results, in eight European countries, 23 March 2009–28 March 2010



Acknowledgements

We are most grateful to Juan Antonio Blasco Amaro for his support to this work, and to Adrian Burton for the translation of the manuscript.

References

1. World Health Organization (WHO). Pandemic (H1N1) 2009 press briefings Audio-visual files and transcripts from the briefings, April 2009. Available from: http://www.who.int/mediacentre/multimedia/swineflupb_20090426/en/index.html
2. EUROfLU. WHO/Europe influenza surveillance. Copenhagen: WHO Regional Office for Europe. [Accessed 1 Apr 2010]. Available from: http://www.euroflu.org/cgi-files/bulletin_v2.cgi
3. Cheng CK, Lau EH, Ip DK, Yeung AS, Ho LM, Cowling BJ. A profile of the online dissemination of national influenza surveillance data. BMC Public Health. 2009;9:339.
4. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012-14.
5. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Goss CH. Does Google Influenza Tracking Correlate with Laboratory Tests Positive for Influenza? Am J Respir Crit Care Med. 2010;181:A2626
6. Google.org. Flu Trends. Google.org. [Accessed 1 April 2010]. Available from: http://www.google.org/flutrends/
7. Wilson N, Mason K, Tobias M, Peacey M, Huang QS, Baker M. Interpreting "Google Flu Trends" data for pandemic H1N1 influenza: The New Zealand experience. Euro Surveill. 2009;14(44):pii=19386. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19386.
8. Kelly H, Grant K. Interim analysis of pandemic influenza (H1N1) 2009 in Australia: surveillance trends, age of infection and effectiveness of seasonal vaccination. Euro Surveill. 2009;14:(31):pii=19288. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19288.
9. Doornik JA. Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data. 8th Oxmetrics User Conference, March 18, 2010. Available from: http://www.gwu.edu/~forcpgm/JurgenDoornik-final-Doornik2009Flu-Jan31.pdf
10. European Centre for Disease Prevention and Control (ECDC). European Influenza Surveillance Network (EISN). Stockholm: ECDC. [Accessed 1 Apr 2010]. Available from: http://www.ecdc.europa.eu/en/activities/surveillance/EISN
11. Réseau Sentinelles France. Situation Epidémiologique en France métropolitaine. Paris. [Accessed 1 Apr 2010]. Available from: http://websenti.b3e.jussieu.fr/sentiweb
12. Sistema de Vigilancia de la Gripe en [Influenza Surveillance System in Spain]. Red Nacional de Vigilancia Epidemiológica. [Accessed 1 Apr 2010]. Available from: http://vgripe.isciii.es/gripe
13. Robert Koch Institute (RKI). Berlin:RKI. [Accessed 1 Apr 2010]. Available from: http://www.rki.de
14. Smittskyddsinstitutet (SMI). Stockholm:SMI. [Accessed 1 Apr 2010]. Available from: http://www.smittskyddsinstitutet.se
15. European Commission. Eurostat. Luxembourg. [Accessed 1 Apr 2010]. Available from: http://epp.eurostat.ec.europa.eu
16. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron AJ. More diseases tracked by using Google Trends. Emerg Infect Dis. 2009;15(8):1327-8.
17. Valdivia A, Monge-Corella S. Diseases tracked by using Google Trends, Spain. Emerg Infect Dis. 2010;16(1):168.