Web query-based surveillance in Sweden during the influenza A(H1N1)2009 pandemic, April 2009 to February 2010

A Hulth (anette.hulth@smi.se)[1], G Rydevik[1]
1. Swedish Institute for Communicable Disease Control, Solna, Sweden

Citation style for this article:
Hulth A, Rydevik G. Web query-based surveillance in Sweden during the influenza A(H1N1)2009 pandemic, April 2009 to February 2010. Euro Surveill. 2011;16(18):pii=19856. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19856

Article published on 5 May 2011

At the Swedish Institute for Communicable Disease Control, statistical models based on queries submitted to a Swedish medical website are used as a complement to the regular influenza surveillance. The models have previously been shown to perform well for seasonal influenza. The purpose of the present study was to evaluate the performance of the statistical models in the context of the influenza A(H1N1)2009 pandemic, a period when many factors, for example the media, could have influenced people's search behaviour on the Internet and consequently the performance of the models. Our evaluation indicates consistent good reliability for the statistical models also during the pandemic. When compared to Google Flu Trends for Sweden, they were at least equivalent in terms of estimating the influenza activity, and even seemed to be more precise in estimating the peak incidence of the influenza pandemic.

Introduction

For this paper, we evaluated the performance of statistical estimates of influenza impact based on queries made on a national medical website. The statistical models were trained on data collected during almost four influenza seasons and were applied to web query data collected during the influenza A(H1N1) pandemic period, since April 2009. Our evaluation concerned both the estimates produced by the web query-based system and their usefulness.

Monitoring of an influenza pandemic relies on a number of surveillance sources. Traditionally, the two main variables collected are the number of laboratory-confirmed cases and the percentage of patients with influenza-like illness among total visits to appointed sentinel general practitioners. These are two standardised influenza surveillance measures recommended by the World Health Organization and the European Centre for Disease Prevention and Control [1,2]. In addition, other sources are used, both formal and informal. In recent years, surveillance based on search behaviour on the Internet has appeared as a potential complement to the traditional sources [3-10]. As the conclusions drawn about the spread and the impact of a pandemic influenza will (or at least should) affect policy makers, it is crucial to evaluate the performance of such additional surveillance methods.

We have previously described a syndromic surveillance system [7] for seasonal influenza which is built on anonymous queries submitted to the search engine of a Swedish medical website: http://www.vardguiden.se. The Vårdguiden website had about 1.2 million visits in January 2010, of which approximately 800,000 were unique. The site is operated by Stockholm county council and around half of the visitors in 2010 originated from the Stockholm region [11] which covers about one fifth of the 9.3 million inhabitants in Sweden. The number of Internet users in Sweden is high: 88% of the population aged 16 to 74 years used the Internet on at least a weekly basis in 2010 [12]. During the first quarter of 2009, 36% of the users in Sweden looked for health-related information on the Internet [13].

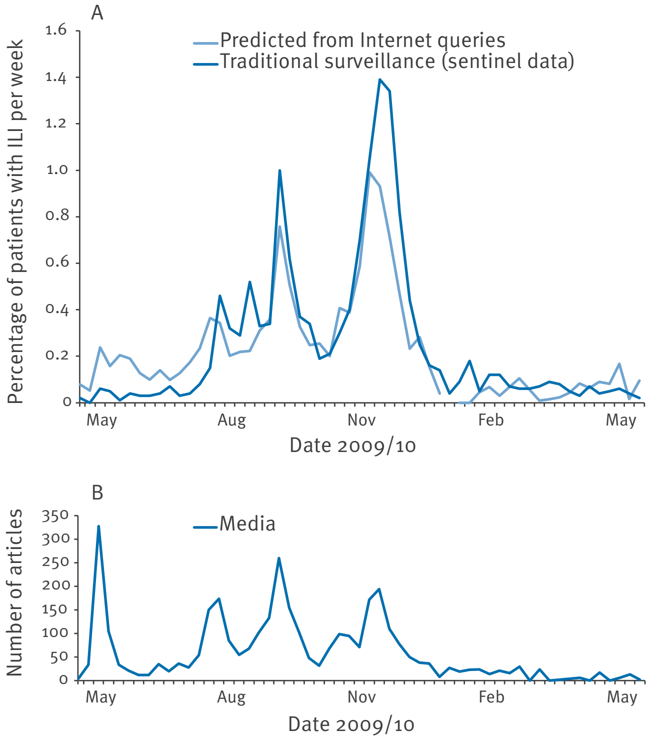
Our statistical models estimate the influenza burden in Sweden [14] and are trained to approximate the number of laboratory-confirmed cases of influenza and the proportion of patients with influenza-like illness reported by sentinel general practitioners. These estimates are based solely on the number of queries about influenza and influenza symptoms (in total 20 types of queries [7]) submitted to the Vårdguiden search engine. The statistical method behind the models has been described in Hulth et al. [7]. The system, which generates a final output in the form of graphs, is fully automatic, including daily transfer of query logs from the medical website to the Swedish Institute for Communicable Disease Control (SMI), statistical calculations, and weekly emails presenting the output of the models that are sent to those in charge of the influenza surveillance at the institute. The email contains two graphs showing the estimated number of laboratory-confirmed cases and the percentage of patients with influenza-like illness from week 16 in 2009 up to the week before the email is sent. An example of what data are contained in the output for the sentinel model is shown in Figure 1. Panel A shows the estimates for the percentage of patients with influenza-like illness calculated from the web queries. Panel B shows the number of media articles in Sweden on influenza, aggregated by week. Because of a reporting delay in the sentinel data, the automatic email can as soon as a week has ended give an estimate of what the traditional system will show only several days later.

In addition to the automatic emails, the graphs are published every week (starting week 36 in 2009) on a publicly available web page [15]. The graphs were also discussed, together with information from a number of other sources, at weekly influenza meetings held at SMI during the most intense phase of pandemic surveillance in 2009/10. By comparing and contrasting the results from the different systems, the epidemiologists got a more complete picture of the spread and the extent of influenza activity in the population.

We have previously shown that the statistical models are able to estimate seasonal influenza [7,16]. The purpose of the presented study was to evaluate the performance of the statistical models in the context of the influenza A(H1N1)2009 pandemic. This was a period during which many factors – for example the media – could have influenced people's Internet search behaviour and consequently the models' performance.

Figure 1

Week-by-week web query-based estimates of the percentage of patients with influenza-like illness among all patients seen, Sweden, week 16, 2009–week 19, 2010



ILI: influenza-like illness.

As three web query logs were missing from week 53 in 2009, the entire week was removed.

Methods

In order to evaluate the performance of the web query-based influenza surveillance system, we performed one qualitative evaluation and two quantitative analyses. The qualitative evaluation consisted of a structured interview with key persons who received the output of the statistical models for use as one source of information on the spread of the influenza pandemic in the country. In the quantitative evaluation, we focused on the performance of the sentinel model, as the traditional laboratory reporting indicated exceptionally high influenza levels, far higher than any of the other surveillance systems and did probably not correctly reflect the influenza impact in Sweden [17]. Here we compared the output from the web query model to the reference data produced by the traditional surveillance, focusing on the potential advantage of the model output with respect to reporting delays in the sentinel data. In a second analysis, we compared our estimates to those made by Google Flu Trends for Sweden [18].

Evaluating the usefulness of the output

An email was sent to five persons at SMI who were deeply involved in the surveillance and the analysis of the spread and the impact of the influenza A(H1N1)2009 virus on the national level. The email contained five questions on the usefulness of the web query-based influenza surveillance, concerning the information conveyed in the graphs as well as the means through which it was distributed. We also asked the users to suggest improvements that could be made to the system. We obtained replies to this email from four persons.

Reporting delays

The sentinel reporting system suffers from reporting delays, since it relies heavily on manual reporting. The reporting delay for the sentinel data for seasonal influenza is up to three weeks during the influenza season, and can be up to five weeks in the beginning of a season [19].

Our statistical model was trained on historical data (week 27 in 2005 to week 15 in 2009) that were back-populated and thus included late reports. As the data in the traditional influenza surveillance are aggregated by week, we chose the same aggregation level for the model based on web queries. The evaluation period covered 44 weeks, from week 16 (13 April) in 2009 to week 6 (14 February) in 2010.

Two quantities were calculated for the statistical model versus sentinel data as reported for a given week (here called 'incomplete sentinel') as well as versus the final sentinel values, including late reports, five weeks later (here called 'complete sentinel'). These quantities were:

1. the root mean squared error of prediction (RMSEP). This is one of the standard measures in model evaluation [20], calculated by

$$\frac{1}{n}\sum(observed - predicted)^2$$

2. the mean absolute deviation (MAD). This value is calculated by

$$\frac{1}{n}\sum|observed - predicted|$$

The advantage of the latter measure is that it is slightly more intuitive than RMSEP, since it tells us how far off the predictions were on average.

We also calculated the R-squared measure, as well as the correlation coefficient.

Comparison with Google Flu Trends

Google Flu Trends was launched for Sweden in October 2009 [18]. In this analysis, we compared the estimates done by Google Flu Trends, which were based on queries submitted from Sweden to the general-purpose search engine Google, to those made by our system based on queries submitted to the national Vårdguiden web site. More specifically, since Google Flu Trends was developed on sentinel data, we used the web query-based sentinel data for the comparison. Both sources aggregate data by week, although Google Flu Trends starts the week with a Sunday, whereas our statistical model starts the week with a Monday.

Results

Usefulness

According to the users of the output produced by the web query-based system, the largest contribution of the graphs was as an additional source and a complement to the traditional surveillance. It was stated that one surveillance system is not enough for getting a true picture, and the more sources point in the same direction, the more reliable is the interpretation of the influenza surveillance data. The automatic dispatch was much appreciated and the emails, sent three and a half days before the time when the traditional surveillance was compiled, was valuable as an early signal of what to expect from the traditional surveillance, although it was the trend rather than the height of the curve that was deemed more important.

As part of the graphs produced by the web query-based system, the crude numbers of articles on influenza in online media (obtained from http://www.eniro.se/nyhetssok/) were plotted. The users appreciated that some indication of the media activity was shown in the graphs. It was, however, evident from the answers that we obtained that some of the users believed that this information was corrected for in the statistical estimates.

Two improvements were suggested to the models: that they should be corrected for the impact of media reports on search behaviour; and that they should be divided into the various regions of the country. This latter wish is, however, impossible to fulfil with this particular data source, as no geographical information is stored in the anonymous query logs. One user requested a better explanation of the model's statistics.

Summary evaluation statistics

In Table 1 we summarise the comparison of our model and Google Flu Trends with the actual sentinel reports. The Swedish sentinel model based on web queries predicted the sentinel numbers better when delayed reporting was taken into account, no matter what performance indicator was used. This makes sense because we trained the models on complete sentinel data. In other words we have, by training the models on data including late reports, obtained a system which better mimics the values we will get after a while, once the data have been back-populated.

The MAD value of 0.15 can be compared with the change from 1.11 percentage points to 1.36 percentage points between week 45 and 46 in 2009 [21,22], during the height of the pandemic. Thus, the average deviation paralleled the weekly change during the most intense pandemic period.

Table 1

Evaluation statistics for models predicting influenza burden based on Internet queries, Sweden, 2009/10

Data	Root mean square error of prediction (percentage points)	Mean average deviation (percentage points)	Coefficient of determination R-squared	Correlation
Vårdguiden model vs incomplete sentinel	0.21	0.15	0.68	0.88
Vårdguiden model vs complete sentinel	0.17	0.12	0.75	0.90
Google Flu Trends vs incomplete sentinel (both normalised)	NA	NA	NA	0.85
Google Flu Trends vs complete sentinel (both normalised)	NA	NA	NA	0.87

NA: not available.

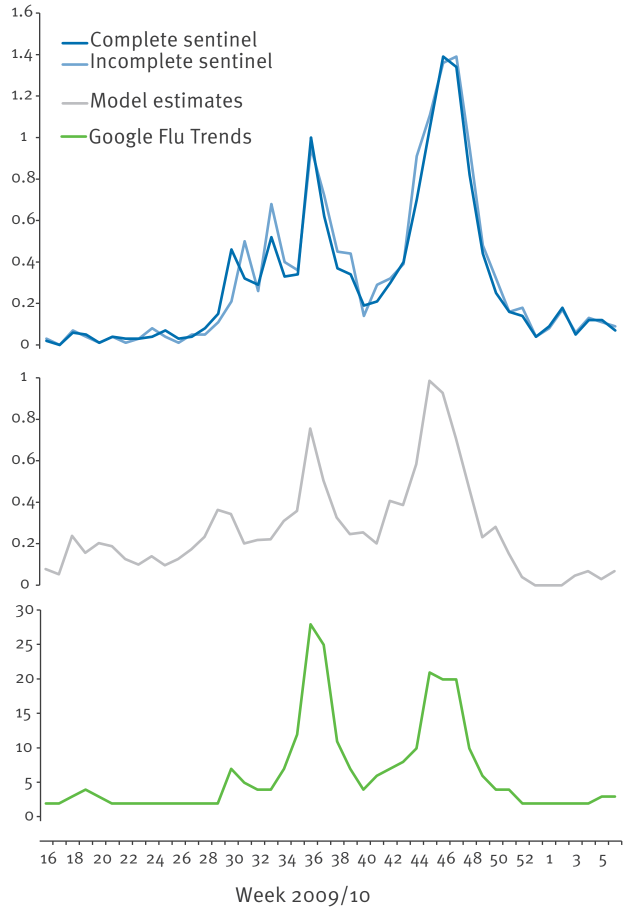
Although the difference is small, the correlation coefficient indicates that our model performed better than Google Flu Trends for Sweden, with a correlation coefficient of 0.90 versus 0.87. Since Google Flu Trends only provides relative intensity indicators and not absolute estimates of reported influenza, R-squared, RMSEP and MAD could not be calculated for their data.

In-depth comparison with Google Flu Trends for Sweden

Figure 2 shows reported sentinel data (incomplete and complete), our web query-based estimations for sentinel data, and Google Flu Trends for Sweden from week 16 in 2009 to week 6 in 2010. When comparing the output of the sentinel model to the traditional surveillance that the model is supposed to mimic (Figures 1 and 2), we can see that the shape of the curves is very similar. The Google data, which are based on more data than the Vårdguiden data, form a smoother curve compared with the output from our statistical model. It underestimates, however, the height of the peak in November 2009.

Figure 2

Incomplete and complete sentinel data, output from the statistical model based on Vårdguiden data, and Google Flu Trends for Sweden, week 16, 2009–week 6, 2010



ILI: influenza-like illness.

Comparison with other published results

The performance of the web query-based sentinel model during the pandemic season in terms of correlation estimates was in line with the performances of various other reported attempts of web-based influenza surveillance (Table 2). The correlation values that have been published are in the region of 0.72-0.94, with one bottom outlier at 0.55 using blog posts [10], and Google's exceptional outlier at 0.96 [6]. The latter is especially surprising given that this value was for correlation with validation data. We found two publications that reported R-square estimates: Eysenbach reported an R-squared value of 0.83 [4], and Polgreen et al. reported an R-squared value of 0.38 [5], but it has to be noted that these were the values obtained when comparing the model to the data used for the fitting process. The R-squared value in our sentinel model denotes the performance relative to previously unknown data, during an exceptional influenza season. In light of this, the estimate of 0.75 is high.

Discussion

Overall, the performance of the statistical models based on queries submitted to the Swedish Vårdguiden web site exceeded our expectations during the pandemic, especially because the models were trained on seasonal influenza. The curve produced by the web query-based sentinel model was very similar to the one obtained from the traditional surveillance the model is supposed to mimic.

We have shown that an independently developed and controlled system such as ours can be comparable in reliability to Google Flu Trends, a model that is trained on much larger data volumes. One downside is that our model has a higher variance, which becomes manifest in numerous small fluctuations of the model estimates in Figure 2, trend shifts that are not reflected in the reported sentinel data. Such false signals can be a cause for concern if the model is to be used to guide public health action, and means in practice that observed trend shifts cannot be trusted unless sustained for two weeks or more.

While others have indicated that the under-estimation of the influenza peak in Sweden of Google Flu Trends could be due to a limitation in the Swedish sentinel system [23], the fact that our model (in addition to other surveillance methods) shows the same pattern as the sentinel reports [17], rather indicates that it is Google Flu Trends that is lacking in the quantitative estimation.

The quantitative evaluation statistics also indicate good reliability. It is debatable, however, whether they are suitable for evaluating surveillance systems for communicable diseases. Such measurements tend to investigate the performance in estimating absolute levels of activity, and give equal weight to the entire period of investigation, including periods of low activity. In future work, it might be more important to look at how a surveillance system captures the dynamics of the disease, such as rapid increases in activity levels or the timing of peaks.

We have also described the results of a qualitative evaluation in which we interviewed four colleagues who were receiving the output from the statistical models. In summary, it was valuable for those working with the surveillance to have an additional source of information, as this increased their confidence in their estimates and predictions of the spread and the impact of the influenza A(H1N1)2009 virus.

One unknown factor here is the media impact on search behaviour. The interviewees explicitly asked for media activity to be incorporated in the statistical model. Such a model should intuitively perform better than a model without this information. We have performed some early experiments on including media activity in our web query-based statistical models. However, we have not yet found a satisfactory model to correct for the assumed impact of media reporting on peoples' search behaviour.

Conclusions

In this paper, we have described an evaluation of a syndromic surveillance system based on queries submitted to the search engine on a Swedish medical website and regularly used during the pandemic influenza period. From our experience, we can say that there are a number of advantages of using web queries as a source for surveillance during a pandemic:

- The system is fully automatic;
- The estimates are produced earlier than the traditional sources that it is supposed to mimic;
- They do not require people to see a doctor;
- There is no reporting delay in the system;
- The system is cheap to maintain;
- A system based on web queries can easily be adapted to different symptoms or diagnoses.

In addition, the presented analyses demonstrated that the system is reliable, stable and performs well when compared with conventional surveillance systems. When comparing the output from our sentinel model to Google Flu Trends for Sweden, we can conclude that although our models had been trained on a substantially smaller set of data, they were at least equivalent to Google Flu Trends in terms of performance, and in terms of peak estimation even seemed to be more precise.

No current method can, however, give us the true spread and impact of an infectious disease in society. Until such a method is invented, the best we can do is to use multiple sources for surveillance, be it an influenza pandemic or another infectious disease. Syndromic surveillance based on web search behaviour clearly has a role to play as such a source.

Acknowledgments

Thanks to Vårdguiden for granting access to the query logs and to Euroling for implementing parts of the automatic transfer. Thanks also to the colleagues at SMI involved in the web query-based influenza pandemic surveillance: Birgitta Brink, Mia Brytting, Annika Linde, and Moa Rehn. We would also like to thank Sharon Kühlmann-Berenzon and the anonymous reviewers for valuable comments on the draft version. Anette Hulth was financed by The Swedish Civil Contingencies Agency.

Table 2

Reported performances of different web-based influenza surveillance systems

Input data	Reported value	Measure	Influenza measure	Reference	Comment
Health information web access logs	0.78, 0.76	Correlation (two different periods)	Sentinel reports, United States	[3]	Values were obtained from calibration data
Regression model using clicks on a sponsored Google Adsense keyword	0.83 0.90	R-squared correlation	Laboratory reported cases, Canada	[4], Figure 2, Table 1	Values were obtained from calibration data
Clicks on a sponsored Google Adsense keyword	0.81	Correlation	Sentinel reports, Canada	[4], Table 1	Value was obtained from calibration data
Regression model using web queries	0.38	R-squared	Sentinel reports, United States	[5]	Value was obtained from calibration data (average R-squared for nine different regions)
Regression model using web queries	0.85	Correlation	Sentinel reports, United States	[6], Figure 2	Value was obtained from calibration data
Regression model using web queries	0.96	Correlation	Sentinel reports, United States	[6], Figure 2	Value was obtained from validation data
Blog posts	0.55	Correlation	Sentinel reports, United States	[10]	Value was obtained from calibration data
Google FluTrends	0.94 (Germany) 0.72 (Poland)	Correlation	Acute respiratory infection (Germany), Influenza-like illness (Poland)	[23]	Values were obtained from validation data (highest and lowest values of all evaluated countries)

References

1. FluNet. [Internet]. Geneva: World Health Organization. [Accessed on 12 Jan 2011]. Available from: http://www.who.int/csr/disease/influenza/influenzanetwork/flunet/en/
2. European Influenza Surveillance Network (EISN). [Internet]. Stockholm: European Centre for Disease Prevention and Control. [Accessed 12 Jan 2011]. Available from: http://www.ecdc.europa.eu/en/activities/surveillance/EISN/Pages/index.aspx
3. Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, Dowling J, et al. Analysis of Web access logs for surveillance of influenza. Stud Health Technol Inform. 2004;107(Pt 2):1202-6.
4. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc. 2006:244-8.
5. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. Clin Infect Dis. 2008;47(11):1443-8.
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):1012-4.
7. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLoS One. 2009;4(2):e4378.
8. Wilson K, Brownstein JS. Early detection of disease outbreaks using the Internet. CMAJ. 2009;180(8):829-31.
9. Brownstein JS, Freifeld CC, Madoff LC. Influenza A (H1N1) virus, 2009--online monitoring. N Engl J Med. 2009;360(21):2156.
10. Corley CD, Cook DJ, Mikler AR, Singh KP. Text and structural data mining of influenza mentions in Web and social media. Int J Environ Res Public Health. 2010;7(2):596-615.
11. Hjort P. Fortsatt höga besökssiffror på Vårdguiden.se. [Continued high numbers of visitors on Vårdguiden.se]. 8 Feb 2010. In: Vårdguiden Labs [Blog]. Swedish. Available from: http://www.vardguidenlabs.se/2010/02/08/fortsatt-hoga-besokssiffror-pa-vardguidense/
12. IT bland individer. [IT usage of individuals]. Stockholm: Statistiska centralbyrån. [Accessed on 12 Jan 2011]. Swedish. Available from: http://www.ssd.scb.se/databaser/makro/produkt.asp?lang=2&produktid=LE0108
13. Privatpersoners användning av datorer och Internet 2009. [Use of computers and the Internet by private persons in 2009]. Stockholm: Statistiska centralbyrån; 2010. Swedish. Available from: http://www.scb.se/Pages/PublishingCalendarViewInfo____259924.aspx?PublObjId=10199
14. Årsrapporter om influensasäsongen. [Annual reports on the influenza season]. The National Influenza Reference Centre (SMI). Stockholm: Smittskyddsinstitutet. Swedish. Available from: http://www.smittskyddsinstitutet.se/publikationer/arsrapporter-och-verksamhetsberattelser/smis-arsrapporter-om-influensasasongen/
15. Webbsök – uppskattningar av influensa baserat på sökningar på vårdguiden.se. [Web search – estimations of influenza based on searches on vårdguiden.se]. Stockholm: Smittskyddsinstitutet; 22 Sep 2010. Swedish. Available from: http://smi.se/publikationer/smis-nyhetsbrev/webbsok/
16. Hulth A, Rydevik G, Linde A. Web Queries for Influenza Monitoring. Poster presentation. European Scientific Conference on Applied Infectious Disease Epidemiology; 19-21 Nov 2008; Berlin, Germany.
17. Influensa A(HINI) 2009 – utvärdering av förberedelser och hantering av pandemin. [Influenza A(H1N1) 2009 – analysis of preparation and management of the pandemic]. Stockholm: Socialstyrelsen; 2011. Swedish. Available from: http://www.socialstyrelsen.se/publikationer2011/2011-3-3/Sidor/default.aspx
18. Explore flu trends – Sweden. [Internet]. Google. Available from: http://www.google.org/flutrends/se
19. Influensarapporter. [influenza reports]. Stockholm: Smittskyddsinstitutet. Swedish. Available from: http://www.smittskyddsinstitutet.se/publikationer/smis-nyhetsbrev/influensarapporter/
20. Lindgren BW. Statistical Theory. London-New York: Chapman and Hall; 1993. 633 pp.
21. Influensarapport vecka 45 (2/11-8/11), 2009. Stockholm: Smittskyddsinstitutet; 12 Nov 2009. Swedish. Available from: http://www.smittskyddsinstitutet.se/publikationer/smis-nyhetsbrev/influensarapporter/sasongen-20092010/influensarapport-vecka-45-2009/
22. Influensarapport vecka 46 (9/11-15/11), 2009. Stockholm: Smittskyddsinstitutet; 19 Nov 2009. Swedish. Available from: http://www.smittskyddsinstitutet.se/publikationer/smis-nyhetsbrev/influensarapporter/sasongen-20092010/influensarapport-vecka46-2009/
23. Valdivia A, Lopez-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 from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19621