Author’s response to reviews

Title: Using internet search queries for infectious disease surveillance: screening diseases for suitability

Authors:

Gabriel J Milinovich (gabriel.milinovich@qut.edu.au)
Simon MR Avril (simon.m.r.avril@gmail.com)
Archie CA Clements (director.rsph@anu.edu.au)
John S Brownstein (john.brownstein@childrens.harvard.edu)
Shilu Tong (s.tong@qut.edu.au)
Wenbiao Hu (w2.hu@qut.edu.au)

Version: 4
Date: 5 December 2014

Author’s response to reviews: see over
Dear Dr Lee,

Please find a revised version of our paper, titled “Using internet search queries for infectious disease surveillance: screening diseases for suitability”, and our response to the reviewers’ comments. Firstly, I would like to thank both the reviewers for providing comprehensive and very useful reviews of our work. We have attempted to address all the comments made. While many of the comments are insightful and preempt future work (and in a number of cases have provided us with ideas for future directions), a number of these are not within the scope of this publication and reasons are provided below as to why some suggestions were not adopted in the revised version.

We thank you for consideration of our paper for publication in the BMC Medicine and look forward to your reply.

Regards

Gabriel Milinovich B.Ag.Sc, MPH, PhD
Postdoctoral Research Fellow
School of Public Health and Social Work
Queensland University of Technology
Kelvin Grove, 4056
Email: gabriel.milinovich@qut.edu.au
**Reviewer:** Cecile Viboud

**Reviewer's report:**

This is a ‘proof of concept’ study exploring the use of Google search queries to monitor a variety of infectious diseases in Australia. Authors assess the correlation between 64 monthly disease time series based on official notifications and Google queries over a 4-year period.

**Major Compulsory Revisions**

1) While I appreciate the systematic application of Google search queries to a variety of diseases, my main issue with paper is that it is rather simplistic. My understanding is that the analysis is limited to univariate correlations with predefined search queries (Table 1, supp table), which is very different from the original Google Flu Trends algorithm that was based on a weighted combination of several indicators. Hence some of the diseases that are shown to have low correlation with Google queries may actually be better monitored by different and more complex search terms. In other words, some multivariate analyses would be welcome. It is possible that I misunderstood the approach, but if so, the methods section and table 1 do need to be clarified.

We agree that the use of multivariate models is a better approach; internet-based surveillance systems reduce the effect of media bias and shifts in internet searching behaviours by incorporating a large number of weighted search terms. This approach was used in the Google Flu Trends model – the revised model used 160 weighted search terms (Cook, Samantha, et al. "Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic." PloS one 6.8 (2011): e23610). The goal of our research, however, was not to replicate Google Flu Trends – this would be a somewhat redundant exercise. Rather, our goal was to develop internet-based surveillance systems for other infectious diseases that increase the capacity for public health units to detect and respond to emerging infectious disease events.

As the reviewer suggested, simple surveillance systems built on a single search term will not suffice. Development of more comprehensive models (such as Google Flu Trends) is both time and resource intensive. It is not possible for us to construct such models for 64 infectious diseases and currently there is insufficient information in the scientific domain for us to effectively prioritise which diseases we should focus our research efforts on. It was for these reasons that we employed the simple methods in this paper. This study has allowed us to identify which disease we should target and lays the
foundation for further investigation into specific diseases. We hope that it will also drive other researchers to develop internet-based surveillance systems for other diseases that have not previously been explored (at the time of submitting this paper we could only find publications describing the use of internet based surveillance for seven of the 65 diseases included in this study, and 5 of the 17 diseases found to be significantly correlated with internet search metrics).

We have made amendments to the stated aims and methods sections of this paper to try to make the justification for and interpretation of the study clearer. Regarding the inclusion of multivariate analyses, we agree that this is the logical next step; however, we feel that this is not within the scope of this publication. We have added more discussion with regards to the limitations of the study and future research directions.

2) Additional measures of synchronicity should be considered; including how well indicators match in terms of peak timing, epidemic onset month, or annual no of cases...
We agree with the reviewer regarding the usefulness of such measures with respect to identification of search terms for use in internet-based surveillance models. We do not, however, believe that this is within the scope of this study (see response to comment 1) and feel that there would be more value in investigating this at a finer resolution (weekly data) and this will form a component of future research on prioritized diseases.
It should, however, be noted that cross correlations do give an indication of synchronicity. The supplementary material has been amended to include all cross correlations for both 2009-13 and 2004-13 data.

3) It would be interesting to highlight the specificity of the correlation results shown in Table 1. Specifically, show that the correlation between disease X (eg, pneumococcal disease) and search term X (pneumonia) is strongest than with any of the other search terms shown here
We understand the reviewers’ point; however, this would create issues owing to the large number of correlations that would need to be performed. The approach used involves 211 correlations between search terms identified for specific diseases (see list of correlation in supplementary material); these were performed for each state and using national Google Trends data to analyse the state notification. For this approach there were potentially 3,587 correlations. When designing this study we considered performing correlations between all identified search terms (164) and diseases (64). Within Australia there are eight states/territories; correlations were performed at state level and against these using national Google Trend data. Using this approach, we would potentially have 178,432 correlations to analyse. Whilst we agree with the reviewer’s
comment that it would be interesting to evaluate the specificity of the correlation results, we believe that such an approach is unnecessary at this point for the reasons discussed in response to point 1 (the primary aim was to identify which diseases to focus on for comprehensive future studies).

We think that it is also prudent to point out that the goal of this study was not to identify the most appropriate search terms to use for inclusion in a surveillance system. This will be a component of future studies that utilize finer resolution data (weekly) and will focus on diseases shown to have most promise in this study, of which pneumococcal is of particular interest. It is our intention to develop a more comprehensive list of potential search terms and look at these in greater detail in future research.

4) State-specific analyses return conflicting results, as compared with national analyses (Table 1). Authors brush on this. What can explain the lack of correlation in some of the state-specific datasets?

As the reviewer points out, there is clearly some spatial variation. It is difficult to identify the exact cause of this from this work. This may be a function of population size, climatic or other factors. The data utilized for this study is not at sufficient temporal or spatial resolution to adequately investigate this further and we intend to include this in future studies on prioritized diseases. Further discussion has been included in the paper.

5) Authors note that in some cases, the correlation between disease notifications and search queries is essentially explained by a common time trend (generally increasing). It would be pretty straightforward to detrend all time series before calculating correlations.

A detrending approach was used for all cross correlations in this study; seasonal differencing was applied. This uses the values of the previous year to remove seasonal effects. The effect of this can be seen in the cross correlations below for the time series “gonococcal infection” and the search term “staph infection” (raw data is presented in figure 1, panel 2). Seasonal differencing has been applied to the right graph, but not the left.
6) Correlations are estimated at the monthly time scale over a 4 year period and hence are based on few observations. It would be good to calculate correlations over longer time periods when available, as a sensitivity analysis (eg, series 1, 3, 15 in Fig 1).

The supplementary tables have been amended to include correlations, p-values and cross correlations for longer time periods (2004 onwards) where available.

Minor Essential Revisions

7) Supplementary table, tab “Google correlate results” was quite scrambled when I downloaded it, perhaps because of international conversion of characters? Perhaps best to save this table as pdf.

A PDF version of this table is provided. I think, however, it is possible that the original table was not scrambled. Google correlate often returns some strange results. For example, the top three results for Hepatitis B (newly acquired) in Column C actually are “å±€å±€å±€”, “å±€å±€å±€å±€” and “å±€å±€å±€”. Regardless, a PDF version has been included with the resubmission.

Discretionary Revisions

8) Are official notifications available by age in Australia? If so, are specific age groups more highly correlated with internet search queries than other?

We agree with the reviewer that this would be an interesting aspect to study (as well as other factors that may enable us to stratify the data – socioeconomic status, gender, rural/urban, etc). The data accessed for this study, however, were publically available, but did not provide such details. The only available data were case numbers (aggregated monthly) and state/territory of origin. Such data are available for Australia; however, gaining access to this is a timely and complicated process (involving ethics from multiple institutions and individual data request approvals for each state and nationally for each disease). It would not be possible to do this for all 64 notifiable diseases in Australia. This is why we decided to undertake a screening study first, to identify which of these diseases we should focus our efforts and resources on. It was also our hope that this study could be used in a similar way by other researchers. Our future research will take these factors into account (for a smaller number of prioritized diseases).
Reviewer: Gerardo Chowell

Reviewer’s report:

Authors analyzed correlations at national and state levels of internet search data and infectious disease surveillance data. Their results are interesting, but are not surprising as a number of studies have already put forward the idea that internet query data have potential to improve epidemiological surveillance. I have some suggestions:

1) I would have liked to see the potential for their simple correlation models to be validated with a subset of their data or additional surveillance data for recent years. At this point, authors only report correlations, which is interesting, but as I said, it is not necessarily novel.

The application of internet-based data to monitoring of infectious diseases has, as the reviewer states, already been put forward by a number of other studies. Prior to this publication, however, the scope of this work was limited; at the time of submitting this paper we could only find previous studies for seven of the 64 diseases analysed in this study. It should also be noted that the methodological approaches differed in previous studies making comparison difficult. Whilst the ultimate goal of our research is to produce robust and accurate models of disease in the community that utilize internet search metrics to produce actionable and timely alerts, our rationale for undertaking this study was to provide baseline data on which to build these models and to determine our research priorities. Prior to undertaking this study, we did not know which diseases would be best to focus our efforts on. We believe our approach is justified in our findings; 12 of the 17 infectious diseases found to be significantly correlated with internet search metrics had not been previously reported in the scientific literature. Our future research efforts will focus on the diseases identified in this study and utilizing an approach that incorporates the suggestions of both reviewers. For example, time series models including seasonal autoregressive integrated moving average, polynomial distributed lag time series models and Bayesian spatiotemporal Conditional Auto-Regressive models will be used to develop an early warning system for these selected infectious diseases. Clearly our research goals were not sufficiently articulated in this document. A number of changes have been made to make the research aims clearer. More details on future research directions have also been added to the discussion.
2) Issues with the use of google flu trends have been recently discussed (see e.g. Lazler et al. Science 2014). This needs to be discussed further and perhaps elaborate in a way forward.

We are aware of this and other publications that have raised similar issues with the use of internet-based surveillance systems (eg. Olson DR et al. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS computational biology 2013 AND Butler D. When Google got flu wrong. Nature 2013). These are clearly issues that need to be resolved. We have added some additional text in the introduction regarding this. We feel, however, that this is largely outside the scope of this document (see our responses to comment 1 from both reviewers).