

The Future of Disease Prevention
Using Google Flu Trends to Detect Outbreaks

Reza Mirza

Psychology: Research Specialist

Reza.mirza@utoronto.ca

The speed at which enterovirus and severe acute respiratory syndrome (SARS) propagated demonstrated Taiwan's need for an even earlier detection system. They adopted the RODS system, which has been effective in providing real-time trend recognition; however, another tool can provide supplement the RODS system and provide an earlier warning system. This tool is Google's flu trends model, which utilizes syndromic surveillance by correlating specific search engine queries with the rate of influenza-like illness (ILI) in the population. The Google model has correlated amazingly with official ILI data and provides faster coverage than any other system available. The model is still in development and in the future it may provide trends for any sort of infectious illness.

Over the past number of years, Taiwan has been the victim of a number of major epidemics. The first epidemic occurred in 1998: the first enterovirus (EV71) outbreak suddenly occurred and killed 78 people. In 2000 and 2001, two smaller outbreaks occurred that killed 25 and 26 people respectively (Lin, Twu, Ho, Chang & Lee, 2003). The second epidemic, which illustrated Taiwan's need for an early detection system, was the SARS outbreak in early 2003. The SARS epidemic began innocently in Taiwan: on April 22nd, 2003 the World Health Organization (WHO) reported 3947 cases and 229 deaths worldwide, whereas Taiwan only had 29 probable cases and zero deaths. However, by June 1st, Taiwan had 680 cases and 81 deaths, which was a six-fold increase of cases in less than a month (Hsieh, Chen & Hsu, 2004). Currently, novel influenza A (H1N1) is potentially a pandemic and in Taiwan it has killed 28 individuals to date. Clearly, after these epidemics – and particularly because Taiwan is not afforded the expertise of the World Health Organization (WHO) because they have been repeatedly denied membership – Taiwan needs to develop its own systems to prevent another epidemic.

The Taiwanese government took appropriate action after the SARS epidemic and invested into the Real-time Outbreak and Disease Surveillance (RODS) in 2004 to prevent another outbreak. The RODS software works by automatically collecting clinical data from hospitals and regions, over-the-counter drug sales and other data in real time (Tsui, F., Espino, J., Dato, V., Gesteland, P., Hutman, J. & Wagner, M., 2003). All the data is then aggregated and parsed to identify any trends that would indicate a potential epidemic. The RODS software can identify specific types of illnesses, based on the symptoms and reported or the drugs bought, and locate which areas are experiencing the outbreak. However, the RODS system provides subpar surveillance for certain diseases, like in the case of influenza A (H1N1). The first limitation of the RODS system is that many individuals may choose to go untreated because symptoms may be mild at first, which is what happened in many cases of H1N1. Many individuals with H1N1 flu had quite mild symptoms and in some cases the flu even occurred without fever. In fact, H1N1 may have been circulating undetected for months in Mexico and other countries. In cases where individuals with the H1N1 flu did go to the doctor, the patients may have been misdiagnosed due to the lack of fever in many cases. In cases where the symptoms were severe and individuals did attempt to receive care, too much time may have passed by the time the trend is noticed by the RODS software. When patients go to hospitals and congregate in waiting rooms, it is difficult to stop transmission unless precautions are taken prior to the arrival of patients: 73% of all traceable SARS infections in Taiwan occurred in a hospital setting (Hsieh, Chen & Hsu, 2004). However, the RODS software would not be able to predict an impending epidemic prior to patients visiting hospitals, since it primarily uses data from those visits to identify the trend.

An even earlier detection system for these infectious outbreaks would allow for rapid response, which could reduce the transmission and impact of the infection. Ginsberg et al. (2009) describe a novel and exciting new method to detect outbreaks early to supplement current detection methods. They examined health-seeking behaviour via search engine queries and found that certain queries were highly correlated with seasonal influenza and influenza outbreaks – including the recent H1N1 strain. The model Ginseng et al. developed shows independence of regional variations in the United States and shows remarkable accuracy. Their model was tested against data provided by the United States Centers for Disease Control and Prevention (CDC) on observed ILI and was shown to have an average correlation of .97 across 42 points (Ginsberg et al., 2009). Moreover, another clear benefit of this model is that it can report current trends with a lag of only one day (and potentially also on a real-time basis), whereas current methods employed by the CDC and the European Influence Surveillance System take 1-2 weeks to report current trends. Similarly, the RODS system in Taiwan requires individuals to either buy medication or go to the hospital, which may not happen in many cases. The advantage this model has over other syndromic surveillance is that people generally first check the Internet, as it is generally far more convenient than visiting a doctor. By having data available prior to the outbreak and the rush of individuals to hospitals, the government can take actions to reduce the spread of illness such as by ordering vaccines; increasing awareness through media reports; preparing hospitals by setting up negative-pressure isolations rooms; quarantining contacts of confirmed cases and taking other necessary precautions.

All the authors of the paper work at Google (except Lynnette Brammer, who works at the CDC) and they used the million of Google search queries each day to create the model. The goal of the model is to predict the percentage of visits to a doctor that are related to ILI. The first step

in developing the model was identifying the top 50 million Google search queries for each week and each American state from 2003-2007. The model uses a single explanatory variable: the probability that a query is correlated with the percentage of doctor visits to ILI, validated by the CDC data. To determine which queries would be used, the authors developed an automated method that would compare each of the top 50 million queries for a week with CDC ILI data. The program preferred queries that demonstrated regional variations that matched the CDC ILI data. They examined the top 100 search queries that correlated with the CDC ILI regional data and found that aggregating the top 45 produced the best fit. When the final model was compared with CDC data for 2003-2007, the mean correlation was .90. The model was then checked against data from 2007-2008 – which was never used to develop the model, unlike the 2003-2007 data – and it had a mean correlation of .97 (min=0.92, max=0.99). Figure 1. is the model's prediction alongside US CDC data, which illustrates the accuracy of the model.

Figure 1: United States Flu Activity – Influenza estimate (Google Flu | Trends How Does this work., 2009).

■ Google Flu Trends estimate ● United States data



Adopting the model and applying it to Taiwan may not be so easy, though. One potential problem with adopting this model in Taiwan specifically is that Google is not the primary search

engine in Taiwan. Thus they cannot provide the service because they do not have access to the majority of search engine queries, which would surely reduce the accuracy of the model. A possible solution would be to contact Yahoo or YAM, the leading search engines in Taiwan, and try to collaborate with either one of them to try to provide their data. Furthermore, this exact model may not be as accurate when applied to other countries, such as Taiwan, as search engine patterns may differ. However, a new model could be developed using the same methodology but instead of inputting American data for regression, the Taiwan Department of Health could provide their own data to input.

Another potential issue with applying this system in Taiwan is regarding differences that exist between the populations of Taiwan and the United States. One key difference is Internet usage rates: in the United States, 72.4% of the population has access to Internet is at 72.4%, which is significantly higher than Taiwan's rate of 65.8% (World Bank, 2009; FIND, 2009). Although Taiwan's Internet usage rate is far lower than that of the States, the Google Flu Trends can still accurately determine infection rates of the entire population by extrapolating from those with Internet access. In fact, the Google Flu Trends has been successfully applied in Australia, where the Internet usage rate is only 55.7% percent of the population (World Bank, 2009). The other population differences include cultural and ecological factors, such as differences in access to healthcare and different mindsets towards Internet usage towards healthcare. The problem arising from these population differences arising is that the Taiwanese may be less likely to search for symptoms than Americans, potentially impairing the effectiveness of the Google Flu Trends system. However, many the Google Flu Trends has been shown to work in many different countries, which include Canada, Spain and Japan. Thus, the Google Flu Trends

demonstrates its versatility and efficacy by accurately capturing flu trends in a number of countries, despite the many population differences that arise.

Currently the Google model by Ginsberg et al. applies only to ILI; however, it is still relatively new and can grow tremendously. The most exciting aspect of this model is that it can potentially provide real-time syndromic surveillance at the first instance of health-seeking behaviour, which is typically searching one's symptoms on the Internet. The model can also identify geographic region through the individual's IP address. By weaving together these two elements, the model can act as a first-response system that detects unusual and sharp increases in queries related to specific symptoms. The model could then analyze the data and map the geographic location of the symptoms and suggest possible explanations based on past trends, or any other data inputted into the system. When it identifies a specific region that is experiencing a surge in the same type of symptoms, it can quickly notify health authorities that there is a potential endemic or form of bioterrorism taking place. Thus, the model affords health officials the ability to mitigate the effects of the transmission by means of preventative measures.

The SARS epidemic was a devastating epidemic in Taiwan and the world at large. In attempt to prevent another outbreak, Taiwan adopted the RODS system. However, this system is limited in its effectiveness due to its method of detection. It relies on visits to doctors and hospital, but the SARS epidemic has clearly demonstrated that once people start visiting the waiting room, transmission is not far behind. Furthermore, the H1N1 outbreak has illustrated another limitation with the RODS system. The H1N1 flu began with mild symptoms and even without a fever, in some cases, which meant many individuals did not consult a doctor. Thus, the RODS system would not be able to detect the outbreak until those individuals that had pre-existing conditions fell ill. The new Google Flu Trend model, developed by Ginsberg et al.

(2009), overcomes these limitations by identifying budding outbreaks before people visit the hospital. By detecting trends at the moment individuals check their symptoms online (which is typically at the onset of the symptoms), health authorities can take the proper measures to prevent epidemics before they even emerge. Furthermore, individuals will check their symptoms online for both the regular flu and the H1N1 flu – as demonstrated by the Google model's correlation with CDC data. The capacity to overcome these limitations coupled with its potential for development suggests that the Google Flu trends model is a noteworthy tool that should be adopted by Taiwan to supplement its current surveillance symptoms.

References

- FIND: Forseeing Innovative New Digiservices. (2009). In Find.org.tw. Retrieved December 20, 2009, from <http://www.find.org.tw/>
- Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, & M. Larry Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012-1014.
- Google Flu | Trends How Does this work?. (2009). In Google.org. Retrieved November 20, 2009, from <http://www.google.org/flutrends/about/how.html>.
- Hsieh, Y., Chen, C. & Hsu, S. (2004). SARS Outbreak, Taiwan, 2003. *Emerging Infectious Diseases*, 10, 201-206.
- Lin, T., Twu, S., Ho, M., Chang, L., & Lee, C. (2003). Enterovirus 71 outbreaks, Taiwan: occurrence and recognition. *Emerging Infectious Diseases*, 9, 291-293.
- Tsui, F., Espino, J., Dato, V., Gesteland, P., Hutman, J. & Wagner, M. (2003). Technical Description of RODS: A Real-time Public Health Surveillance System. *The Journal of the American Medical Informatics Association*, 10, 399-408.
- World Bank Development Indicators. (2009). In WorldBank.org. Retrieved December 22, 2009, from <http://datafinder.worldbank.org/indicators>.