Google Technology in the Surveillance of Hand Foot Mouth Disease in Asia

Rachael Cayce[1*], Kathleen Hesterman[2], and Paul Bergstresser[3]

[1]*West Los Angeles, 5971 Venice Blvd, Los Angeles, CA 90034;* [2]*Tulane University, 1430 Tulane Ave, New Orleans, LA 70112 and* [3]*Department of Dermatology, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd, Dallas, TX 75390-9069, USA*

Abstract: Hand Foot Mouth Disease (HFMD) is a worldwide Enteroviral infection. Severe outbreaks have recently occurred in the US and Asia. Google technology has been shown to predict influenza epidemics and is a potential resource to track epidemics in developed countries where the use of Web-based searches is prevalent.

Google Trends and Google Correlate were used to enumerate Web-based search queries related to HFMD in three Asian regions and were compared to known seasonal variations and standard surveillance data to investigate correlations. We also tested whether a mathematical model, constructed using Google Correlate, would have predicted an outbreak of HFMD in Singapore.

HFMD-related search queries strongly correlated to known HFMD seasonal variation and standard surveillance data. Our mathematical model of Singaporean HFMD predicted the magnitude and chronology of the summer 2012 HFMD outbreak. Google technology shows enormous potential for HFMD surveillance. Further prospective studies are needed to validate the utility of Google technology in HFMD surveillance.

Keywords: Hand foot and mouth disease; Infectious disease surveillance; Google trends; Google correlate; Epidemiology.

INTRODUCTION

Hand, Foot, and Mouth Disease (HFMD) is a worldwide illness caused by Enteroviral infection. Patients commonly present with fever, malaise, and pharyngitis days before onset of a characteristic vesicular eruption on the palms and soles, and posterior pharyngeal erosions. Outbreaks are most common in the summer and generally affect children younger than five years of age [1]. The typical patient recovers with only symptomatic treatment, but recently there have been outbreaks associated with severe disease, such as aseptic meningitis, encephalitis, and even death [2]. Several of these severe outbreaks have been associated with Enterovirus 71 infection in East and Southeast Asia. In the United States, Coxsackie virus A6 was associated with severe HFMD presentations during a large outbreak in the winter of 2011-2012 [2]. Given the lack of specific treatment and vaccination, endemic regions rely on surveillance and public control measures to decrease the frequency of case severity in at-risk populations [3].

The Internet is a frequent source of health-related information for consumers with patients performing Web-based searches for medical information using their symptoms. Google Flu Trends® (GFT), which uses multiple Web-search queries for influenza symptoms to provide real-time influenza surveillance, not only correlates with Centers for Disease Control and Prevention (CDC) surveillance data, it can predict peaks in influenza outbreaks one week faster than CDC data [4]. A related technology, Google Trends® (GT), is a free, online resource that displays the relative frequency pattern or "trend" of Web-searches over time [5]. GT has shown correlation with CDC data for West Nile Virus outbreaks and to previously established patterns of Respiratory Syncytial Virus outbreaks [5]. GT is a potential resource to track epidemics in developed countries where the use of Web-based searches is prevalent. This real-time surveillance capability may identify disease outbreaks early, thereby allowing rapid implementation of control measures to reduce case transmission and population morbidity and mortality.

GT relies on user-generated search queries to produce trends and, for purposes of correlation, requires careful selection of queries. Google Correlate® (GC), which is similar to GFT in its methodology, provides an automated method for query

*Address corresponding to Rachael Cayce, MD, Kaiser Permanente, West Los Angeles, 5971 Venice Blvd, Los Angeles, CA 90034, USA; RachaelCayce@gmail.com; Tel: (323) 857-3713; Fax: (323) 857-2019

selection without the need for conjecture. It also provides modeling capabilities, which have been used for various financial interests, but have yet to be used for infectious disease modeling.HFMD is common in developed countries [2]. It displays strong seasonal variation, epidemic potential, and easily recognizable manifestations, which are ideal features for "trending" using GT and GC. Real-time surveillance of HFMD is especially important given the recent upswing in severe outbreaks in Asia. This report investigates the correlation of HFMD-related search queries, generated using GT and GC, compared to established seasonal patterns and standard surveillance data for HFMD in three Asian regions. We also developed a mathematical model for HFMD in Singapore, constructed using GC, to determine whether this model would, in retrospect, predict the magnitude and chronology of an HFMD outbreak in Singapore in 2012.

METHODS

We outlined epidemiological sources available for HFMD in Asia. We described how GT data are processed, how queries are refined, additional features available using GT, and described how GT trend data were generated for HFMD in Asia for correlation to standard surveillance data. We then described how GC queries are processed, how query searches are refined, additional features available using GC, and described how GC was used to generate correlated HFMD-related search queries in two Asian regions using standard surveillance data. Finally, we described how GC was used to construct a mathematical model of HFMD in Singapore and how this model tested against an outbreak of HFMD.

Google Trends (GT, http://google.com/trends/)

GT generates index graphs based on a user's Google search queries. It displays a relative search volume index graph based on a fraction of total Google Web searches over a specified time period and extrapolates the data to estimate total search volume. This information is updated daily.

All query outputs are normalized by dividing data sets by the largest variable in the data set to allow for comparison of variables within a data set. To increase sensitivity for detection of changes in future trends, GT also divides by an unrelated, and common, Web-search query. Normalization also factors out the effect of a larger population on search volume trends, making it possible to rank cities.

GT scales the result for each query entry relative to its average search volume over the period selected by dividing each data point by the highest query value and multiplying by 100 to produce data on a scale from 0 to 100.

The use of symbols can sophisticate query searches. For example, an addition sign (+) combines searches whereas a subtraction sign (-) excludes searches. A quotation mark will restrict the query to Google searches with words in that order. Query filters comprised of dates, locations, search types, and categories are applied to limit the data. Users logged into a Google email (Gmail) account may download the query results into a comma separated value (CSV) file.

Individual GT HFMD data sets were gathered for three Asian regions. Search terms used the predominant language(s) in each region. In Singapore, the predominant languages of English and Chinese were chosen due to a lack of data generated using Tamil and Malay terms for searches. Initial search terms incorporated the name of the illness (Hand Foot Mouth, Hand Foot Mouth Disease), the symptoms of the illness (hand rash + foot rash + mouth rash) and the viral etiology (Enterovirus). The search term 'hand and mouth' was omitted to avoid data for 'hoof and mouth disease', which affects cloven-hoofed animals. Searches for the symptoms and Enterovirus did not generate query data and were excluded from analysis. GT data were either limited (China) or not available (Macao, Korea) for HFMD and, thus, these regions were excluded from analysis. Results were downloaded into CSV files for statistical analysis.

The Emerging Disease Surveillance and Response Unit of the World Health Organization (WHO) operates in the Western Pacific region to collect data for HFMD diagnoses within each member state's surveillance system. Epidemiol-

ogical data are available for Hong Kong, China, Macao, Japan, Korea, and Singapore [6].

To determine correlations between GT index graphs and standard surveillance methods, we used (a) visual comparison by placing corresponding data sets side-by-side, and (b) Pearson correlation coefficients (R). We sought to determine whether GT data would strongly correlate (R > 0·75) to surveillance data in Japan, Hong Kong, and Singapore. WHO surveillance system data were accessed to obtain, if available, accurate case counts for comparison to GT and GC data in Japan, Hong Kong, and Singapore.

In Hong Kong, HFMD was not a notifiable disease for the Centre for Health Protection (CHP), which collected weekly case counts from sentinel monitoring clinics (64 general outpatient clinics and 50 private general practitioners) and hospital authorities [7]. CHP provided the rate of HFMD consultations per 1000 total consultations and the number of HFMD-related hospitalizations. From these data, we compared the weekly cases of HFMD in 2012 to weekly GT query values (Figure **1**).

In Japan, HFMD was not a notifiable disease and the National Institute of Infectious Disease (NIID) collected weekly case counts for HFMD from pre-fectural public health institutes and 464 sentinel monitoring clinics [8]. In 2012, the rate of HFMD in Japan was relatively low, and therefore, we chose to extend the study period from January 2010 to April 2013. Since the NIID published less than one year of data, we used the WHO records, which were available since 2010 and were based on the NIID data collection. From these data, we calculated the average monthly case counts from January 2010-April 2013 to compare to monthly GT query values (Figure **2**).

In Singapore, HFMD was a notifiable disease, and the Ministry of Health (MOH) collected weekly case counts from sentinel clinics, general practitioners, and childcare centers [9]. From these data, we compared the weekly cases of HFMD in 2012 to weekly GT query values (Figure **3**).

## Google Correlate (GC, http://www.google.com/trends/correlate/)

In contrast to GT, which supplies a search frequency over time for entered query(ies), GC supplies search queries with similar frequency patterns to user-generated search queries, user-provided data sets uploaded to GC, or user-provider drawings of a frequency pattern.

Hong Kong
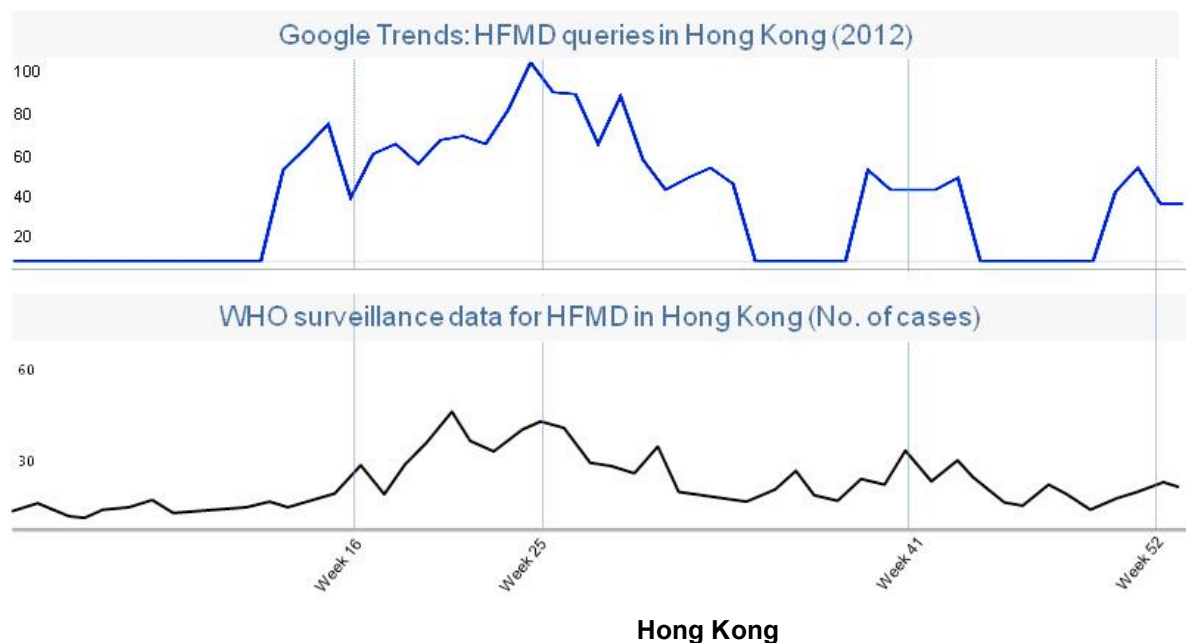
Figure 1: GT query (top) for HFM + HFMD (手足口病 + 手足口) compared to WHO epidemiological curve (bottom) for HFMD in 2012. "100" represents peak search interest in GT Web-searches. Weekly time points for peaks are highlighted for comparison between the two graphs.
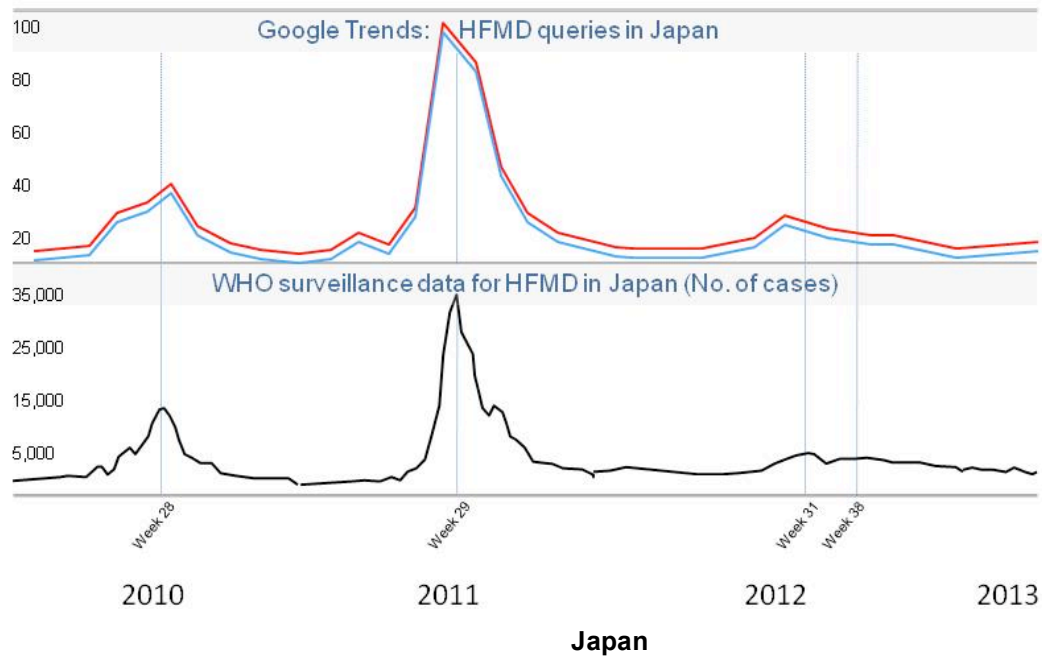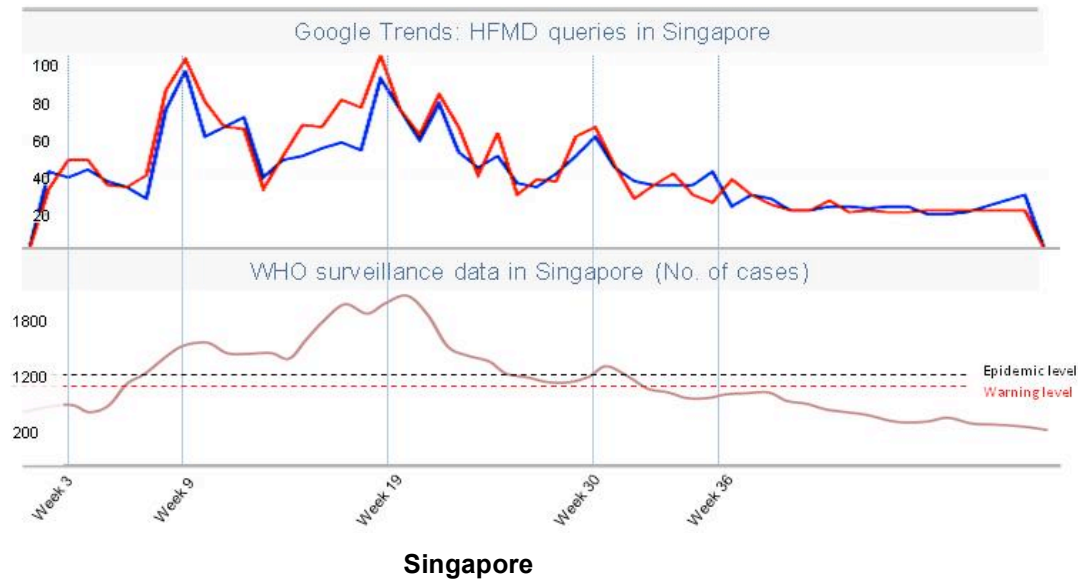
Figure 2A: GT query (top) for HFM + HFMD (手足口病, 手足, red and blue, respectively) compared to WHO epidemiological curve (bottom) for HFMD in Japan between January 2010 and April 2013. "100" represents peak search interest in GT Web-searches. Weekly time points for peaks are highlighted for comparison between the two graphs.
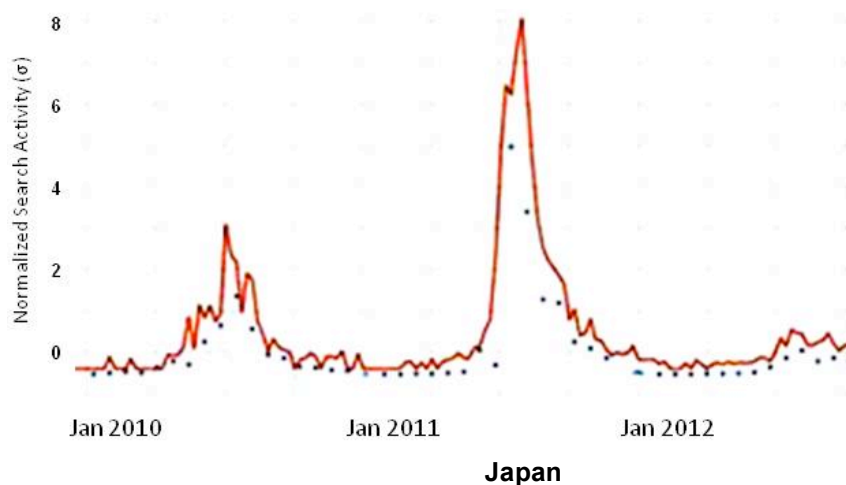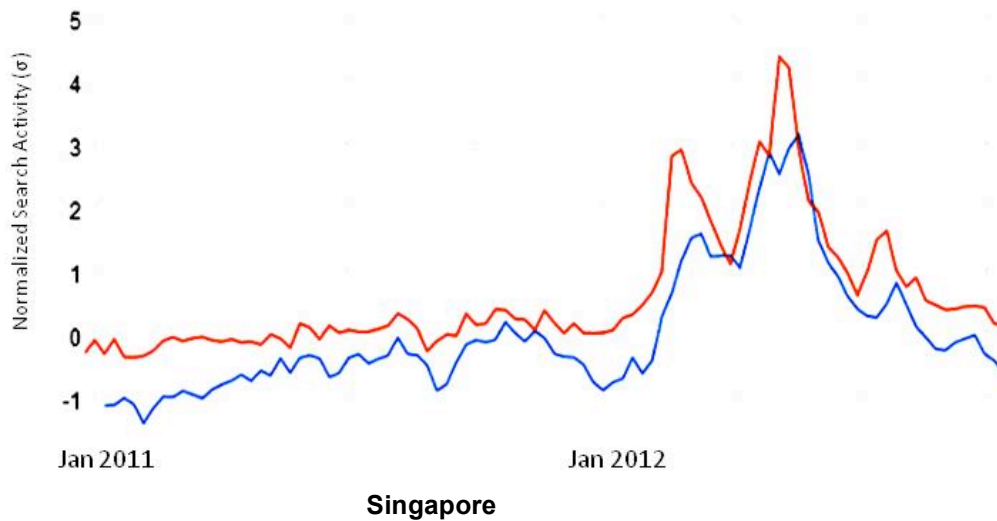

Figure 2B: GC line chart generated using WHO HFMD surveillance data in Japan between January 2010 and December 2012 (blue dots) plotted against the most correlated search query [HFMD infection (手足 口 病 感染, red line].

GC employs a two-pass system to achieve efficient (99% precision) data comparisons between target and GC database series. It produces results similar to that of the GFT's batch-based approach, but at faster rates and with fewer computational requirements. As a result of this analysis, Pearson correlation computation was performed between target and GC database series for exact correlation.

All GC database time-series were normalized by dividing by the total count for all queries in that week and, thus, were controlled for annual growth in Web-based searches. Each time-series was standardized to have a mean value of zero and a variance of one so that queries could be compared using Pearson correlations. GC output included line charts and scatter plots that visually presented correlated times-series. Users were

Singapore

Figure 3A: GT query (top) for HFM+ HFMD (手足口+ 手足口病, blue and red, respectively) compared to WHO epidemiological curve (bottom) for HFMD in Singapore 2012. "100" represents peak search interest in GT Websearches. Weekly time points for peaks are highlighted for comparison between the two graphs.

Singapore

Figure 3B: GC line chart generated using Singaporean MOH HFMD surveillance data in in 2012 (blue) plotted against the most correlated search query [HFMD, red].

allowed to categorize correlations against space (comparing within US States) or time (comparing weekly or monthly time-series). GC allowed time-series to be shifted to determine whether shifting of the series increased correlation with a specific time-series in the GC database. GC also had a 'Location' filter to select for time-series data within specific countries.

Surveillance data sets for Singapore and Japan were tabulated into CSV files and uploaded to GC. In Singapore, weekly case counts, provided

by the MOH from January 2011 to December 2012, were used for analysis. In Japan, average monthly case counts provided by the WHO from January 2010 to December 2012 were used for analysis. Hong Kong was excluded because it was not a location option on GC. GC-generated output queries were catalogued for their relationship to HFMD and we sought to determine whether HFMD search-queries would correlate strongly (R > 0·75) with frequency patterns for Japanese and Singaporean HFMD surveillance data.

Using GC, query results were combined to generate a mathematical model of the data of interest [10]. Singapore was chosen for model generation due to the availability of accurate weekly data points. To build a mathematical model of HFMD in Singapore, Singaporean HFMD surveillance data from January 2011 to February 2012 (termed surveillance training data) and September to October 2012 were uploaded to GC as is standard for model "training". This surveillance training data generated a large set of correlated queries to the data set. The 20 most correlated queries (correlation coefficients [R] = 0.9247 - 0.864) were downloaded into a CSV file for model generation and labeled as "query training data" (Table **1**). As seen in the Table, the highest correlated queries logically related to HFMD, but other GC-generated queries only related temporally to HFMD. For example, "Of the Gods" references to the Nine Emperor Gods Festival, which occurs annually in late summer or early fall and, thus, correlated with peak outbreaks of HFMD. The linear mathematical model used for the study had the following formula:

Table 1: Query Inputs Used to Generate HFMD-Related GT Index Graphs

1.	hfmd
2.	Hand foot and mouth
3.	Hand foot
4.	Mouth disease
5.	Foot and
6.	Hand foot and mouth disease
7.	Foot and mouth
8.	Foot and mouth disease
9.	Know lyrics
10.	Of the gods
11.	Fondant
12.	The gods
13.	å±‹
14.	Feast
15.	I used
16.	I used to
17.	Foot mouth
18.	Hand foot mouth
19.	lv handbags
20.	Rooftop

Table Caption: Top 20 GC-generated correlated queries to Singaporean HFMD surveillance data for January 2011 to February 2012 and September to October 2012.

Model = a × (sum of queries) + b

Stdev = standard deviation

a = surveillance training data stdev/query training data stdev

b = surveillance training data average - surveillance training data stdev × query training data average/query training data stdev

Based on the training data, the resulting linear model was tested against the Singaporean surveillance data for January 2011-October 2012 to determine if the model correlated with the summer outbreak of HFMD in 2012 (March to August 2012) in timing and severity (Figures **4A** and **4B**).

RESULTS

Hong Kong: A statistical analysis between Hong Kong CHP sentinel clinic surveillance data and GT query for HFMD showed a moderate correlation (R=0.64, p=.0000003) (Figure **1**). An analysis between Hong Kong CHP hospitalization surveillance data and GT query for HFMD showed a strong correlation (R=0.78, p<0.000001), data not shown. Several peaks for GT data appeared to precede a peak in standard reporting (weeks 16, 41, and 52).

Japan: Searches for HFM + HFMD (,) were largely superimposed and searches for HFMD were used for analysis against WHO surveillance data (Figure **2A**). A statistical analysis between the WHO surveillance data and GT query for HFMD showed a strong correlation (R=0.97, *p<0.0000001*).

The GC-calculated correlation between the query for "HFMD infection" and the WHO surveillance data was strong (R=0.9485, *p<0.0000001*) (Figure **2B**).

Singapore: A statistical analysis between the Singapore MOH surveillance data and GT query for HFMD showed a strong correlation (R=0.86, *p=<0.0000001*) (Figure **3A**). Several peaks in the GT data appeared to precede peaks in the standard report (weeks 19, 30, and 36).

The GC-calculated correlation between query for HFMD and the Singaporean MOH surveillance data was strong (R=0.947, *p<0.0000001*) (Figure
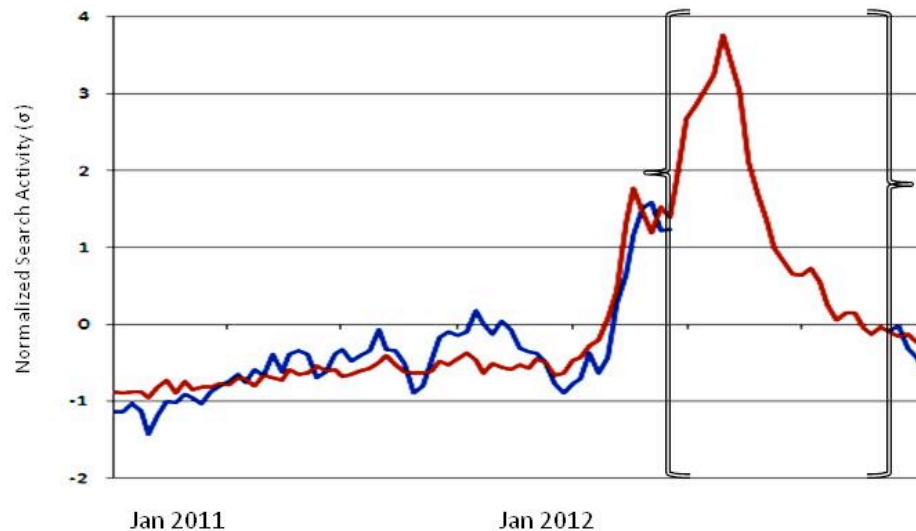
Singapore's GC-based Linear Mathematical Model Generated Using MOH Surveillance Data

Figure 4A: GC-based linear mathematical model (red) generated using Singapore's MOH surveillance data for HFMD (blue) for January 2011-February 2012 and September-October 2012. Surveillance data withheld from model construction is indicated using brackets.
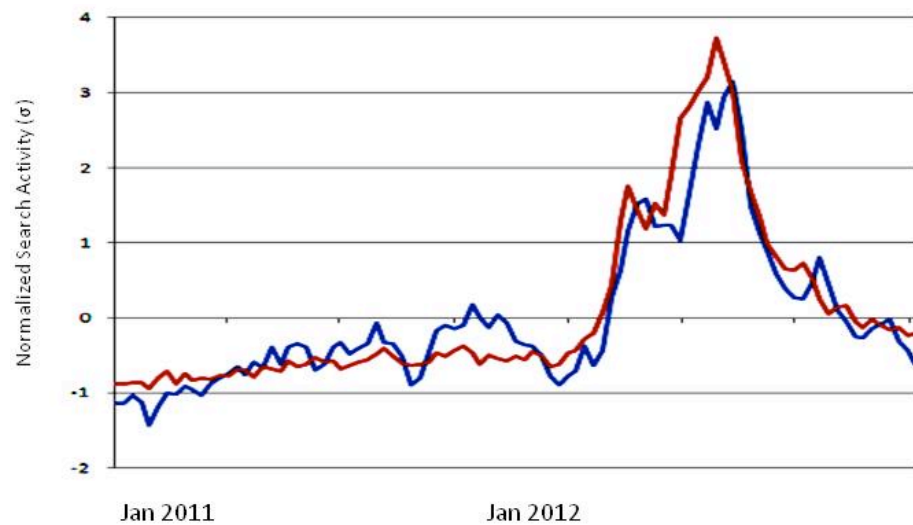
Singapore's GC-based Linear Mathematical Model Generated Using MOH Surveillance Data

Figure 4B: The GC model (red) was able to predict the Singaporean summer (March-August) outbreak of 2012 as depicted by the full Singaporean surveillance data (blue).

3B). Queries for "Hand Foot Mouth" and "Hand Foot Mouth disease" also strongly correlated to Singaporean MOH surveillance data (R=0.9429 and 0.9376, respectively) (data not shown).

A GC-based linear mathematical model of Singaporean HFMD predicted the outbreak of HFMD in the summer of 2012 (Figure **4B**). The model appeared to overestimate the magnitude of the outbreak, but was able to mimic, if not precede, the chronology of the outbreak.

DISCUSSION

Concerns for emerging infectious diseases and bioterrorism heightens the need for real-time surveillance of infectious outbreaks to allow for the rapid institution of public control measures to reduce the frequency and mortality from epidemic illnesses, including HFMD [3]. Various social media are currently being explored, such as Twitter, Facebook and Google, to determine whether their data can be used to develop accurate models of

infectious disease. GT is well situated for this purpose, given its worldwide use, existing technology, and demonstrated correlation with standard surveillance data for several infectious diseases [5]. This ability is highlighted by the fact that GFT can predict influenza outbreaks one week faster than conventional reporting methods from the CDC [4].

The GT queries generated for HFMD correlated with established seasonal variations and surveillance data in three Asian regions. GT index graphs for Singapore and Japan strongly correlated with HFMD surveillance data, and in Hong Kong and Singapore the GT peaks appeared to precede the peaks in standard reporting. In Hong Kong, GT index graphs did not strongly correlate with surveillance data, but the reporting methods from the Hong Kong CHP were limited for epidemiological comparisons. The CHP did not adjust the weekly consultation rate for the population and, therefore the ability to make incidence comparisons was limited. Furthermore, the CHP provided hospitalization counts on a weekly basis that were likely to have underestimated the actual case count for HFMD since most cases were mild and did not require hospitalization. Enteroviral 71 infection was a notifiable illness in Hong Kong, but not all infections manifest symptoms, particularly HFMD, making these data very limited for comparison to HFMD-related query data. Using GC, HFMD-related search queries strongly correlated with surveillance data in Japan and Singapore. In the case of Singapore, our GC-based model was able to predict, possibly at a faster rate, real world surveillance data. Therefore, Google-dependent technology may be useful for disease surveillance in the regions of Asia most at risk for HFMD.

GT and GC may improve disease surveillance, but in their current forms, are most useful when combined with conventional surveillance methods. GT search criteria are not standardized. Search terms need to be validated using the model of GFT or "syndromic surveillance". "Syndromic surveillance", used in electronic medical records, standardizes search terms into syndromes that are validated against case data [5]. GC was developed to overcome this particular limitation of GT by providing correlated queries to data of interest.

However, GC provides a limited set of correlated queries and, for some data sets, the list may contain unrelated queries. Furthermore, limitations to GT and GC include limited or unavailable trending data in regions at risk for HFMD epidemics, such as Macao, Korea, China, and Hong Kong (in the case of GC). Perhaps most importantly, demographic information about Internet-users and the relative volume of searches by healthcare workers, as compared to patients, is unknown by researchers using GT and GC. Finally, our GC-based model appeared to overstate the magnitude of the outbreak, which was similar to the GFT predictions for the 2012 influenza outbreak in the United States [11]. This suggests that Internet-based models may be most useful for real-time surveillance and less useful for predictions of case counts during an outbreak. This potential for real-time surveillance, however, makes Internet-based modeling a potentially useful tool for infectious disease monitoring where rapid implementation of public control measures would be critical to reduce population morbidity and mortality.

Our results using Google technology correlate strongly with known HFMD seasonal variations and standard surveillance data in three Asian regions. Additional real-time analyses comparing the severity and timing of a HFMD outbreak to predictions made by GT and, especially, GC are needed to confirm the usefulness of Google technology in Asian HFMD surveillance where large outbreaks of HFMD, with fatalities, are common. Further, case interviews with Web users searching for HFMD symptoms to determine the specific search terms used and the chronology of such searches in an outbreak are likely to provide powerful data to epidemiologists.

REFERENCES

[1]   Mancini A, Shani-Adir A. Other Viral Diseases. In: Bolognia JL, Jorizzo JL, Rapini RP, editors. Dermatology. 2nd ed. Maryland Heights, MO: Mosby Publishing; 2003. p. 1220.

[2]   Hand, Foot, and Mouth Disease [homepage on the Internet]. Atlanta, GA: Centers for Disease Control and Prevention; c2013. [updated 2013 Mar 7; cited 2014 Mar 18]: Available from: http://www.cdc.gov/hand-foot-mouth/

[3]   Ruan F, Yang T, Ma H, *et al*. Risk factors for Hand, Foot, and Mouth Disease and Herpangina

and the preventive effect of hand-washing. Pediatrics 2011; 127: e898-e904.

[4] Ginsberg J, Mohebbi MH, Patel RS, Brammer I, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009; 457: 1012-4.

[5] Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009; 49: 1557-64.

[6] Hand, Foot, and Mouth Disease (HFMD) Epidemiological Curve. World Health Organization, Western Pacific Region. Emerging Disease Surveillance and Response Unit. c2013. [updated 2013 Mar 7; cited 2014 Mar 18]: Available from: *http://www.wpro.who.int/emerging_diseases/ HFMD.epicurves/en/index.html*

[7] Centre for Health Protection [homepage on the Internet]. Hong Kong: Centre for Health Protection Department of Health The Government of the Hong Kong Special Administrative Region.; c2012. [updated 2013 Mar 7; cited 2014 Mar 18]: Available from: *http://www.chp.gov.hk/en/.*

[8] National Institute of Infectious Disease [homepage on the Internet]. Tokyo, Japan: National Institute of Infectious Diseases.; c2013. [updated 2013 Mar 7; cited 2014 Mar 18]: Available from: *http://www.nih.go.jp/niid/en/*

[9] Ministry of Health Singapore [homepage on the Internet]. Singapore: Ministry of Health.; c2013. [updated 2013 Mar 7; cited 2014 Mar 18]: Available from: *http://www.moh.gov.sg/content/moh_ web/home.html*

[10] Google Correlate [homepage on the Internet]. Mountain View, CA: Google Correlate; c2011. [updated 2013 Mar 7; cited 2014 Mar 18]: Available from: http://www.google.com/trends/correlate

[11] Butler D. When Google got flu wrong. Nature 2013; 494: 155-6.