Disease Tracking in GCC Region Using Arabic Language Tweets

Muhammad Usman Ilyas*
Faculty of Computing and Information Technology
University of Jeddah
Jeddah, Mecca Province, Saudi Arabia
milyas@uj.edu.sa

Jalal Suliman Alowibdi
Faculty of Computing and Information Technology
University of Jeddah
Jeddah, Mecca Province, Saudi Arabia
jalowibdi@uj.edu.sa

ABSTRACT

Several prior studies have demonstrated the possibility of tracking the outbreak and spread of diseases using public tweets and other social media platforms. However, almost all such prior studies were restricted to geographically filtered English language tweets only. This study is the first to attempt a similar approach for Arabic language tweets originating from the Gulf Cooperation Council (GCC) countries. We obtained a list of commonly occurring diseases in the region from the Saudi Ministry of Health. We used both the English disease names as well as their Arabic translations to filter the stream of tweets. We acquired old tweets for a period spanning 29 months. All tweets were geographically filtered for the Middle East and the list of disease names in both English and Arabic languages. We observed that only a small fraction of tweets were in English, demonstrating that prior approaches to disease tracking relying on English language features are less effective for this region. We also demonstrate how Arabic language tweets can be used rather effectively to track the spread of some infectious diseases in the region. We verified our approach by demonstrating that a high degree of correlation between the occurrence of MERS-Coronavirus cases and Arabic language tweets on the disease. We also show that infectious diseases generating fewer tweets and non-infectious diseases do not exhibit the same high correlation. We also verify the usefulness of tracking cases using Twitter mentions by comparing against a ground truth data set of MERS-CoV cases obtained from the Saudi Ministry of Health.

CCS CONCEPTS

• **Information systems → Decision support systems**; **Data analytics**; **Online analytical processing**; **Web mining**;

KEYWORDS

Arabic; disease tracking; epidemiology; gulf region; Twitter

ACM Reference Format:
Muhammad Usman Ilyas and Jalal Suliman Alowibdi. 2018. Disease Tracking in GCC Region Using Arabic Language Tweets. In *The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3186357

[*] Also with Department of Electrical Engineering, School of Electrical Engineering and Computer Science (SEECS)National University of Sciences and Technology (NUST), H-12, Islamabad – 44000, Pakistan. Email: usman.ilyas@seecs.edu.pk.

WWW'18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04..
https://doi.org/10.1145/3184558.3186357

1 INTRODUCTION

1.1 Motivation

Tracking the outbreak of diseases is an essential function of state health departments. Departments collect data from hospitals and compile it to produce a comprehensive picture. The typical time between the emergence of a cluster of patients, to its identification and any public announcements is approximately two weeks. In recent years there have been several attempts to leverage the near-realtime availability of web data sources to reduce this lead time. Most of these studies and experiments were conducted on English language content, with others in French [4], Spanish [18] and Chinese [16]. Many of these prior studies were also limited to specific countries or geographic regions. In this study we apply the concept to the Arabian peninsula, which is home to the Gulf Cooperation Countries (GCC, which include Saudi Arabia, Kuwait, Bahrain, Qatar, the UAE and Oman) and Yemen. Moreover, most social media use in the region occurs in the local language, Arabic, with only a fraction in English.

1.2 Problem Statement

The research question we try to answer in this study is the following: *Are techniques for disease tracking from English language web-data equally applicable to Arabic language social media posts from the GCC / Arabian peninsula region?*

1.3 State-of-the-Art

The approaches used in the prior state-of-the-art span three approaches: (1) Collect data from hospitals and health facilities, compile it and detect any disease outbreaks [7, 11, 13, 15]. Since this approach relies on first-hand data this is considered the most reliable and error free approach. However, it comes at the cost of a high lead time of approximately two weeks between the beginning of an outbreak and its detection. (2) Periodically crawl public web data sources, RSS feeds and search engine trend information for reports on disease outbreaks [2, 3, 5, 14, 18]. These approaches are not as reliable as the first approach and are therefore used to compliment slower but more reliable methods based on first hand data. (3) The third type of approach scavenges posts on various social media platforms to detect disease outbreaks [1, 4, 6, 16, 17]. Given the unverified nature of social media posts, these approaches are perhaps least trustworthy. However, they operate on near real-time data which means they have the lowest lag-time.

Moreover, web-based approaches, whether they operate on RSS feeds or social media posts, are language and region dependent.

1.4 Proposed Approach

In this study we collected a data set of Arabic language tweets containing at least one of a number of disease names and that are geo-filtered for the Arabian peninsula. The data sets, one for each disease, span a period of two years and five months. We produce the histogram of the volume of tweets mentioning each disease and observe which diseases appear more frequently on Twitter and see how it relates to the official data.

1.5 Key Contributions

The contributions of this study are two-fold:

(1) Determines which diseases from the list of commonly occurring diseases in the Arabian peninsula are represented in the public Twitter feed of the region.
(2) Determine whether spikes in the mentions of diseases in Arabic-language tweets from the region correspond to spikes in the number of cases reported.

2 RELATED WORK

Brownstein et al. [2] were among the early users of web-sourced data for the early detection of disease outbreaks. They developed HealthMap a website that crawls and retrieves news articles from public news aggregators every 15 minutes.

Another widely used approach leverages the fact that a large proportion of people in the developed world often turn to the Internet to search for advice on symptoms and treatments [2, 3, 5, 14, 18]. Access to search engine query trends are an additional source of near real-time data that can be leveraged. Every search engine query has a timestamp and is tied to a user's IP address, which can be mapped back to the user's location with a significant degree of accuracy. This means search trends can be broken down to a fine temporal and spatial resolution. Microsoft's Bing Keyword Research [12] and Google's Google Trends [8] are two examples of widely used search engines that have made their search query information available to the public. Pelat et al. [14] used this approach to track diseases from Google Trends. Valdivia and Monge-Corella [18] also used Google Trends to validate this approach for Spain and using Spanish language search terms. While the high degree of location and temporal accuracy may make this approach seem ideal it has a serious drawback - it is just as easy for a malicious user to tamper with these trends by deliberately inflating search queries for targeted terms. Butler [3] reported on a flu season in which Google trends drastically overestimated the spread of flu.

Search query based approaches also require can only be relied on in societies that enjoy a high rate of Internet connectivity. In poorer, less well-connected societies like Pakistan disease mapping is done by replacing the dependence on search trends with survey teams and footwork, as documented by Rojahn [15]. Obviously, this approach adds a significant cost that is not incurred in search trend based approaches.

Corley et al. [5] used spinn3r.com, a now defunct blog indexing service, to search blogs for mentions of influenza. They were able to show high correlation between mentions of influenza on spinn3r.com indexed blogs and CDC surveillance data. The study was limited to English language content only was limited to a relatively short period of only 20 weeks.

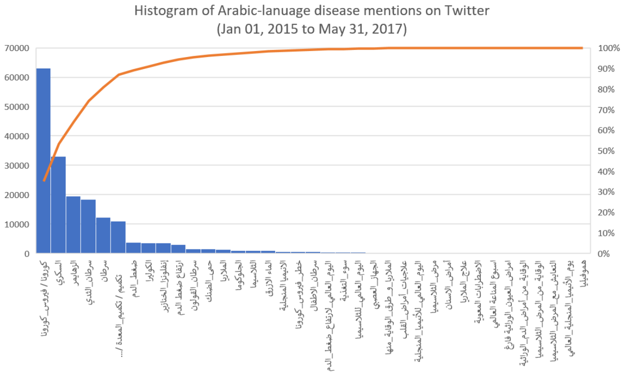


Figure 1: Histogram and cumulative histogram curve of number of unique tweets mentioning each disease.

Culotta [6] took the same approach and applied it to Twitter. However, it too was limited to English language content about influenza cases in the US over just 10 weeks.

Chunara et al. [4] compared the delays between cholera case reports in the aftermath of the Haiti earthquake of 2010. They considered data spanning a period of 100 days from the Haitian Ministry of Health, HealthMap and Twitter. The study was limited to English and French language Twitter feeds.

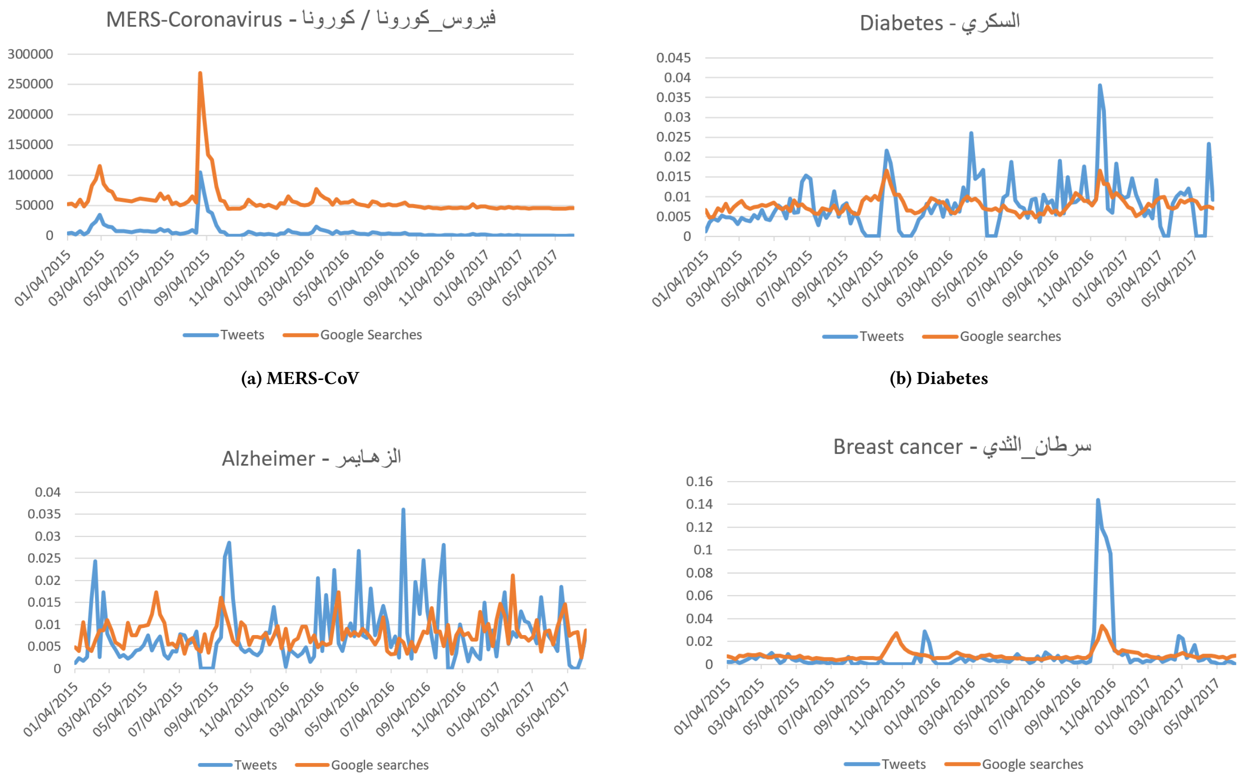
Subsequent surveys of early detection of diseases such as Schmidt [17] and Bernardo et al. [1] have shown that almost prior studies relied on non-Arabic language data and none have been tested and applied to the Arabian peninsula.

3 TWEETS DATA SET

To obtain a list of commonly occurring diseases in the region, we turned to the Ministry of Health of the Kingdom of Saudi Arabia as our reference source [9]. The list is quite extensive and was mapped to the 40 Arabic language keywords shown on the horizontal axis of Figure 1. The tweets were geo-filtered for the Arabian peninsula and span a period of two years and five months, from Jan 01, 2015 to 31 May, 2017. The data set comprises of 180, 503 tweets in total collected using the facilities of the Web Observatory of the King Abdulaziz University in Jeddah, Saudi Arabia. Figure 1 shows a histogram of search terms and the corresponding number of unique tweets that were captured as part of the data set. Also overlaid with the histogram is the cumulative histogram curve. The cumulative histogram curve shows that only the nine most frequently occurring keywords account for 95% of all tweets in the data set. Tweets relating to the Middle East Respiratory Syndrome-Corona Virus (MERS-CoV) are the most common and alone make up for approximately 35% of all tweets in the data set.

The English translations of the nine most commonly occurring disease keywords accounting for approximately 95% of the data according to the histogram in Figure 1 are:

(1) MERS-Coronavirus
(2) Diabetes
(3) Alzheimer
(4) Breast cancer
(5) Cancer



(a) MERS-CoV

(b) Diabetes

(c) Alzheimer

(d) Breast cancer

Figure 2: Time-series plots of tweets and Google searches of various diseases from Jan 01, 2015 to May 31, 2017 in the GCC region.

(6) Sleeve gastrectomy + gastrectomy
(7) Blood pressure + hypertension
(8) Cholera
(9) Swine flu

However, even within this group only the first six contain more than 10% (or 18, 000 tweets) of the complete data set.

4 DATA ANALYSIS

4.1 Tweets and Google Searches

We plotted the time-series of tweets for each disease alongside the volume of searches obtained from Google trend in Figure 2 and Figure 3[1]. Please, note that the number of tweets plotted against each week on the time-axis include the original tweets, their retweet count and their favorite count[2]. Visual inspection shows remarkable similarity between the time-series of Twitter and Google searches for MERS-CoV. This degree of similarity is also quantified by the correlation coefficients listed in Table 1. The high degree of similarity between the time-series for MERS-CoV is borne out by its correlation coefficient of 0.999855. As the entries in Table 1 show, there is a general downward trend in those values, *i.e.* correlation coefficient goes down with decrease in data set size. With the exception of MERS-Cov and Cholera, none of the diseases at the top of the list are infectious or exhibit any seasonality. This is reflected by the noisy nature of their time-series and their correspondingly low correlation coefficients.

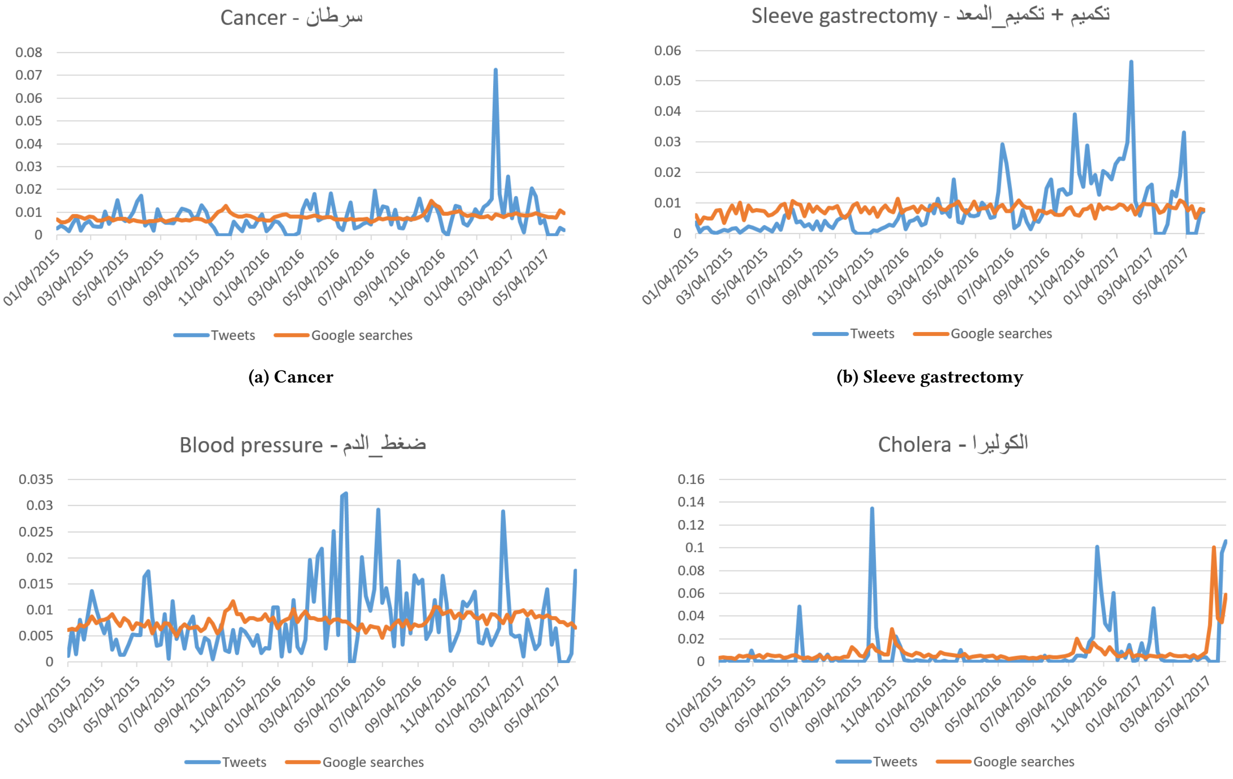
An exception to this rule is breast cancer, which shows two coinciding peaks in both tweet and Google search time-series. These are explained by the fact that October is breast cancer awareness month [10], which are the times when peaks occur in both signals.

Cholera and swine flu, the only other infectious diseases besides MERS-CoV, have relatively low correlation coefficients of 0.316445 and 0.201707, respectively. This is explained by the fact that the data sets for cholera and swine flu are both very small. As a matter of fact, there is a steep drop-off in the size of disease data sets after breast cancer, the fourth on the list.

4.2 Tweets and Case Reports

We continue the analysis of mentions of diseases in Arabic language tweets only for MERS-CoV. Of the other eight diseases hardly any garners as much traffic on Twitter as MERS-CoV. Furthermore, most of them are chronic in nature, rather than seasonal of infectious.

[1] Due to lack of space we have plotted time-series of only the top eight diseases here.
[2] Therefore, the sum of all tweet counts in the time-series plots for a disease will not correspond with the tweet counts in the histogram in Figure 1, which lists only unique tweets



(a) Cancer

(b) Sleeve gastrectomy

(c) Blood pressure

(d) Cholera

Figure 3: Time-series plots of tweets and Google searches of various diseases from Jan 01, 2015 to May 31, 2017 in the GCC region.

In this section we analyzed the time-series of Arabic tweets mentioning MERS-CoV by comparing it with the epicurve[3] of MERS-CoV obtained from the Saudi Ministry of Health [11]. The epicurve data can be considered ground truth for the number of cases. Figure 4 is a plot of the MERS-CoV epicurve and its tweets time-series. Both curves exhibit strong similarities with overlapping peaks. The correlation coefficient of both curves is 0.74, which can be interpreted as modest-to-high correlation The high correlation with the Google search time-series and the moderate-to-high correlation with the epicurve go to show that the time-series of Arabic language tweets is effective at tracking the size of outbreaks of MERS-CoV cases in the GCC region.

5 CONCLUSION

In this study we considered the possibility of monitoring the volume of tweets as a means to detecting outbreaks of diseases in the GCC region. Since the overwhelming majority of Twitter traffic in the GCC region is in Arabic, we used Arabic language keywords. We performed this study for all diseases prevalent in the region, infectious or otherwise, and in the first stage used data from Google trends as ground truth data.

Table 1: Correlation coefficients of tweets with Google searches time-series.

Rank	Disease name	Correlation
1.	MERS-Coronavirus	0.999855
2.	Diabetes	0.420736
3.	Alzheimer	0.044911
4.	Breast cancer	0.678157
5.	Cancer	0.057949
6.	Sleeve gastrectomy	0.160969
7.	Blood pressure / hypertension	-0.059175
8.	Cholera	0.316445
9.	Swine flu	0.201707

We obtained a list of diseases common to the GCC region from the Saudi Ministry of Health. Our analysis shows that only a small subset of those diseases are talkd about on Twitter. The tweets for the top nine diseases account for 95% of the data set we collected over a 29 month period. Of those nine diseases only three (MERS-CoV, cholera, swine flu) are infectious diseases, but only MERS-CoV appears to regularly generate sufficient traffic to be useful for analysis. We also observed that seasonal public awareness campaigns for

[3] An epicurve is a time-series of disease cases.

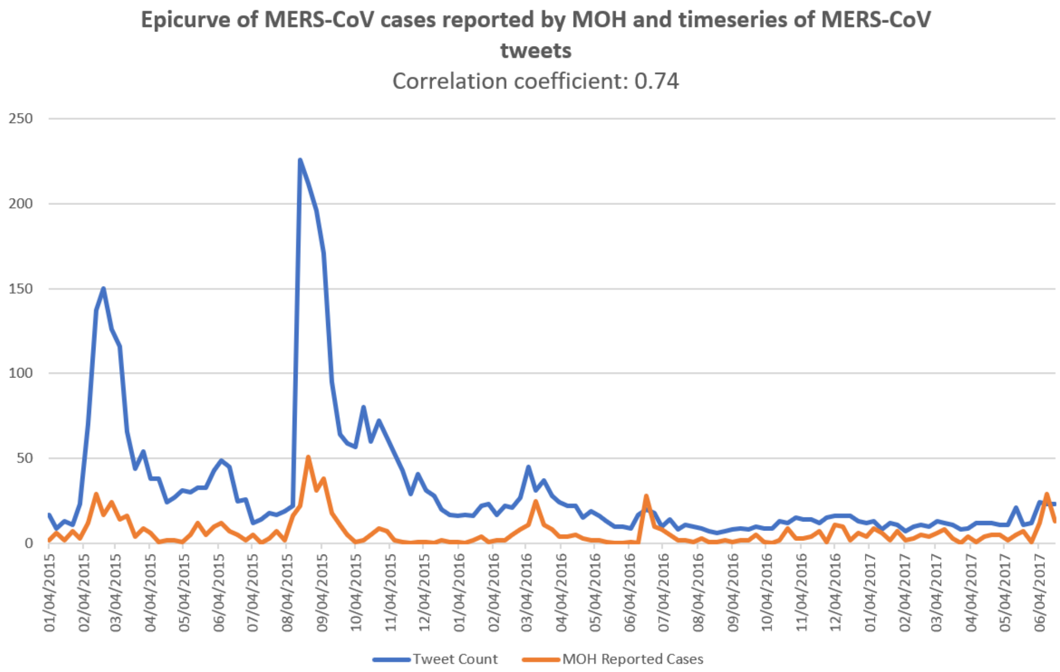


Figure 4: Total tweets and epicurve of MERS-CoV cases.

non-infectious diseases like breast cancer can also exhibit moderate, albeit artificial, correlation.

In order to further establish the correctness of our conclusion regarding MERS-CoV we also obtained the epicurve for it from the Saudi Ministry of Health. Analysis showed that the time-series of the number of tweets about MERS-CoV exhibits a modest-to-high correlation with the epicurve.

We conclude that, at present Twitter traffic volume, in the GCC region MERS-CoV is the only infectious disease for which Twitter traffic may be used for the early detection of outbreaks.

REFERENCES

[1] Theresa Marie Bernardo, Andrijana Rajic, Ian Young, Katie Robiadek, Mai T Pham, and Julie A Funk. 2013. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of medical Internet research* 15, 7 (2013), e147.

[2] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. 2009. Digital disease detectionâĂŤharnessing the Web for public health surveillance. *New England Journal of Medicine* 360, 21 (2009), 2153–2157.

[3] Declan Butler. 2013. When Google got flu wrong. *Nature* 494, 7436 (2013), 155.

[4] Rumi Chunara, Jason R Andrews, and John S Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene* 86, 1 (2012), 39–45.

[5] Courtney D Corley, Diane J Cook, Armin R Mikler, and Karan P Singh. 2010. Using Web and social media for influenza surveillance. In *Advances in Computational Biology*. Springer, 559–564.

[6] Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*. ACM, 115–122.

[7] Center for Disease Control. last accessed: January 15, 2018. FastStats - Infectious Disease. *https://www.cdc.gov/nchs/fastats/infectious-disease.htm* (last accessed: January 15, 2018).

[8] Google. last accessed 18 June 2017. Google Trends. *https://trends.google.com/trends/* (last accessed 18 June 2017).

[9] Ministry of Health Kingdom of Saudi Arabia. last accessed June 10, 2017. Diseases - Diseases List. *http://www.moh.gov.sa/en/HealthAwareness/EducationalContent/Diseases/Pages/* (last accessed June 10, 2017).

[10] Ministry of Health Kingdom of Saudi Arabia. last accessed June 10, 2017. Health Days 2016 - Breast Cancer Awareness Month. *http://www.moh.gov.sa/en/HealthAwareness/healthDay/2016/Pages/HealthDay-2016-10-01-31.aspx* (last accessed June 10, 2017).

[11] Ministry of Health Kingdom of Saudi Arabia. last accessed November 14, 2017. Statistics - Statistics. *https://www.moh.gov.sa/en/ccc/pressreleases/pages/default.aspx?PageIndex=1* (last accessed November 14, 2017).

[12] Microsoft. last accessed 18 June 2017. Bing - Keyword Research. *https://www.bing.com/toolbox/keywords* (last accessed 18 June 2017).

[13] World Health Organization. 2018. WHO | Global Health Observatory (GHO) Data. *http://www.who.int/gho/en/* (2018).

[14] Camille Pelat, Clement Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. 2009. More diseases tracked by using Google Trends. *Emerging infectious diseases* 15, 8 (2009), 1327–8.

[15] Susan Young Rojahn. 2012. Pakistan Uses Smartphone Data to Head Off Dengue Outbreak. *MIT Technology Review* (10 2012).

[16] Marcel Salathé, Clark C Freifeld, Sumiko R Mekaru, Anna F Tomasulo, and John S Brownstein. 2013. Influenza A (H7N9) and the importance of digital epidemiology. *The New England journal of medicine* 369, 5 (2013), 401.

[17] Charles W. Schmidt. 2012. Using social media to predict and track disease outbreaks. *Environmental health perspectives* 120, 1 (2012), A31.

[18] Antonio Valdivia and Susana Monge-Corella. 2010. Diseases tracked by using Google trends, Spain. *Emerging infectious diseases* 16, 1 (2010), 168.