Chapter 28
Using Non-traditional Data Sources for Near Real-Time Estimation of Transmission Dynamics in the Hepatitis-E Outbreak in Namibia, 2017–2018

Michael Morley, Maiamuna S. Majumder, Tony Gallanis, and Joseph Wilson

Abstract *Background:* Google Trends (GT) is an emerging source of data that can be used to predict, detect, and track infectious disease outbreaks. GT cumulative search volume data has been shown to correlate with cumulative case counts and to produce basic and observed reproduction number estimates analogous to those derived from more traditional epidemiological data sources. An outbreak of Hepatitis-E (Hep-E) occurred in Namibia in the fall and winter of 2017–2018. We used GT data to estimate transmission dynamics of the outbreak and compared these results with those estimated via data from HealthMap, a relatively new digital data source, and with surveillance reports from the government of Namibia published in the World Health Organization Bulletin, which is a traditional data source. *Objective:* Aim 1: To determine the correlation between GT relative search volume data (RSV) and cumulative case counts from the HealthMap (HM) and World Health Organization (WHO) data sources. Aim 2: To estimate and compare transmission dynamics including basic reproduction numbers (R_0), observed reproduction numbers (R_{obs}), and final outbreak size (I_{max}) for each of the three sources of data. *Methods*: GT relative search volume data regarding the term "hepatitis" in Namibia was acquired from October 13, 2017–March 2, 2018. Cumulative reported case counts were obtained from the

Michael Morley and Maia Majumder contributed equally and are co-first authors.

M. Morley (✉)
Harvard Medical School, Ophthalmic Consultants of Boston, Boston, MA, USA
e-mail: mgmorley@eyeboston.com

M. S. Majumder
Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Engineering Systems Division, Massachusetts Institute of Technology, Cambridge, MA, USA

T. Gallanis
Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

J. Wilson
Department of Global Health Policy and Management, the Heller School, Brandeis University, Waltham, MA, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_28

HealthMap and WHO data sources. The Incidence Decay and Exponential Adjustment (IDEA) model was used to calculate R_0, R_{obs}, and final outbreak size for the three data sources. *Results*: The correlation coefficient between GT cumulative relative search volume and both HM and WHO cumulative case counts measured R = 0.93. The mean R_0 and R_{obs} estimates for the hepatitis-E outbreak in Namibia were similar between the GT, HM, and WHO data sources and are similar to previously published Hep-E R_0 estimates from Uganda. Final outbreak size was similar between HM and WHO data sources; however, estimates using GT-derived data sources were smaller. *Conclusions*: GT cumulative search volume correlated with cumulative case counts from the HM and WHO data sources. Mean R_0 and R_{obs} values were similar among the data sources considered. GT-derived final outbreak size was smaller than both HM and WHO estimates due to diminishing search volume later in the epidemic possibly due to search fatigue; nevertheless, this data source was useful in describing the transmission dynamics of the outbreak including correlation with case counts and reproduction numbers.

Keywords Hepatitis · Hepatitis-E · Hepatitis-E virus · Google trends · HealthMap · Digital disease surveillance · Media events · Correlation · Reproduction number

Learning Objectives

(1) Access and analyze non-traditional data sources for outbreak surveillance in a low-resource setting.
(2) Model the transmission dynamics associated with an outbreak in a low-resource setting.

28.1 Introduction

28.1.1 Google Trends and HealthMap Data

Google Trends (GT) allows users to obtain search volume data on specified search terms from defined locations and specified time frames (Nuti et al. 2014; Google Trends 2018). GT analyzes a statistical sampling of the 3.5 billion daily Google searches and provides graphical and downloadable data that can be analyzed for many purposes, including public health. The initial enthusiasm over GT's ability to detect and predict infectious disease—namely, influenza like illnesses—was tempered by estimation failures during the 2009–10 H1N1 pandemic, though researchers have since made modifications such that have improved its accuracy and reliability (Yang et al. 2015). GT has been found to be a valuable data source in evaluating infectious diseases occurring in low- and middle-income countries such as malaria in Thailand (Ocampo et al. 2013) and dengue in Bolivia, Brazil, and India (Yang et al. 2017). In addition, GT has been used successfully as a surveillance tool to detect and predict

infectious outbreaks such as influenza, dengue, Zika, and Ebola (Alicino et al. 2015; Yang et al. 2017; Majumder et al. 2016). A growing number of epidemiologic studies show correlation between GT cumulative search volume and cumulative case counts in acute infectious outbreaks. HealthMap utilizes disparate online sources including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports to describe the current global state of infectious diseases and their effect on human and animal health (HealthMap 2019).

28.1.2 Hepatitis-E in Namibia

Hepatitis E (Hep-E) occurs in Namibia at a low baseline rate with periodic outbreaks. The most recent Hep-E outbreak in Namibia occurred in the fall of 2017, among residents living in informal settlements near the capital city of Windhoek. Hep-E infections occur primarily through the ingestion of food or water contaminated with infected feces. Public health risk factors for contracting Hep-E include low economic status, crowded living conditions, inadequate sanitation facilities, and lack of reliable, safe drinking water and food. All of these factors were present in the 2017 Hep-E outbreak in Namibia (2018a). The incubation time of the Hepatitis-E virus (HEV) is 4–6 weeks, and the majority of affected people are asymptomatic or minimally affected making timely detection of active viral shedders difficult (Center for Disease Control and Prevention 2018; World Health Organization 2018b). Behavioral risk factors include open air defecation without toilets and consumption of street food. Both behaviors are noted in Namibia's informal settlements, a term used to describe housing areas used by inhabitants with low socioeconomic status. Environmental risk factors include rainy season (Nov-March in Namibia) during which untreated surface water may be collected and ingested or used for agriculture and other purposes.

28.1.3 Response to Hepatitis-E Outbreak in Fall 2017

A coordinated, multifaceted response to the fall 2017 Hep-E outbreak in informal settlements near Windhoek was organized by the Namibian government with support from the World Health Organization (WHO), United Nations International Children's Fund (UNICEF), United Nations Population Fund (UNFPA), and the Namibian Red Cross (2018a). Multiple approaches were employed to combat the epidemic. These efforts included a campaign to disseminate information to the public regarding the disease and ways to minimize transmission, creation of improved sanitation/toilet facilities and water sources, hand washing awareness, advisories for pregnant women, and water disinfecting tablets. Public communication via newspaper articles, radio announcements, television stories, and social media were used to inform the public. Campaigns to inform the public and community leaders via meetings and forums

were initiated. The Minister of Health and the President of Namibia made public visits to the affected areas reinforcing the messages of sanitation, hygiene, and clean water.

This chapter aims to analyze the transmission dynamics associated with the 2017 Hep-E outbreak in Namibia using Google Trends relative search volume (GT) and HealthMap (HM) data. Results from analyses using the Incidence Decay and Exponential Adjustment (IDEA) model (Fisman et al. 2013) using these non-traditional data sources which are then validated against surveillance reports from the government of Namibia published in the World Health Organization Bulletin, a traditional data source (World Health Organization 2018c). Finally, the utility of non-traditional data sources for infectious disease surveillance in low-resource settings is discussed.

28.2 Methods

28.2.1 Data Sources

Raw epidemiologic data about the 2017 Hep-E outbreak in Namibia was obtained from two sources. Cumulative reported case counts of HEV infections in Namibia were obtained from World Health Organization's (WHO) publicly available bulletins (World health Organization 2018c) released weekly during the time period of this study. This information was collected by WHO and the government of Namibia during the outbreak and is considered the "ground truth". The second source of raw cumulative reported case counts was obtained from HealthMap digital disease surveillance system (HealthMap 2019). HealthMap data includes information automatically collected on-line by algorithms from newspapers, journal articles, bulletins from relief agencies, reports from outbreak monitoring groups, and other sources from which suspected case reports are gleaned. Linear smoothing was conducted to adjust the shape of the HealthMap cumulative case curve using Google Trends search data (GT + HM). No human experimentation was performed, and all work was conducted in accordance with the Helsinki Declaration (1964).

28.2.2 Google Trends

Google Trends relative search volume data regarding the Hep-E outbreak was collected on October 1, 2018 for the dates October 13, 2017–March 2, 2018. GT RSV search data for the search term "hepatitis" was downloaded into Excel and analyzed.

28.2.3 Data Analysis

Using linear smoothing, the cumulative Google Trends relative search volume data was normalized to the HM cumulative incidence curve. The scaling constant was obtained by dividing the HM total cumulative case count (893) by the Google search fraction sum (1493) for the dates October 13, 2017 to March 2, 2018 resulting in a normalization factor of 0.62. By multiplying the cumulative Google Trends relative search volume data by this scaling constant, a third estimate for cumulative Hep-E cases was obtained (i.e. GT + HM).

The weekly search volume for the 68 days prior to the December 20, 2017 peak (i.e., October 13, 2017–December 19, 2017) and the 72 days following the peak search volume (i.e., December 21, 2017–March 2, 2018) were tabulated, as were the number of days with zero searches.

We tracked the number of cases as measured by WHO and HM in a running total (cumulative) format. Correlation between the WHO, HM-only, and GT + HM data was measured using the Pearson correlation coefficient, R.

Finally, WHO, HM, and GT + HM data sources were used to calculate estimates for the transmission dynamics including the mean basic (R_0) and observed (R_{obs}) reproduction numbers as well as the final outbreak size (I_{max}) associated with the Hep-E outbreak in Namibia using the Incidence Decay and Exponential Adjustment (IDEA) model (Fisman et al. 2013). Generalized Reduced Gradient (GRG) non-linear optimization and a serial interval length of 5–9 days was used to parameterize the model. Linear interpolation was used to accommodate missingness across all data sources.

28.2.4 Statistical Analysis

Statistical analysis was performed using Excel. Pearson correlation coefficient, R, was used to measure correlation between WHO and HM cumulative case counts with GT-HM cumulative relative search volume. The basic and observed reproduction numbers are presented as mean, minimum, and maximum.

28.3 Results

28.3.1 WHO, HM, and GT + HM Data

Cumulative case counts from WHO and HM data sources along with normalized cumulative Google RSV data are shown in Fig. 28.1.

The correlation coefficients, R, between GT + HM, HM, and WHO cumulative curves are listed in Table 28.1.

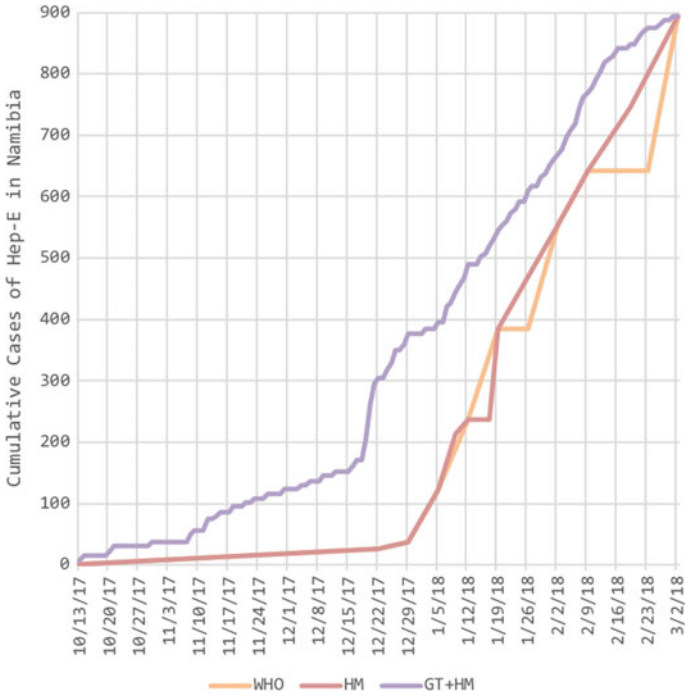

Fig. 28.1 Cumulative Hep-E incidence as ascertained from the WHO, HM, and GT + HM data

Table 28.1 Pearson correlation coefficients across data sources. GT + HM = HealthMap data smoothed with Google Trends relative search volume data; HM = HealthMap data; WHO = World Health Organization data

	WHO	HM	GT + HM
GT + HM	0.935	0.930	1
HM	1	1	0.930
WHO	1	1	0.935

The Google Trends relative search data for the term "hepatitis" in Namibia during the time frame October 13, 2017—March 3, 2018 demonstrated a strong peak on December 20, 2017, and the search volume remained elevated for the next 2 months (Fig. 28.2). The Namibian Government and the WHO sponsored a "media day" to alert the general public and the medical community about the hepatitis-E outbreak on December 20, 2018 and additional public events and interventions occurred during late December 2017 through January 2018 (World Health Organization 2018a).

Table 28.2 describes the increase in the GT relative search volume following the media day event on December 20, 2017. The sum of GT search volume, % of fractions, the number of non-zero search days, and the percent of non-zero search days all increased (Table 28.2).

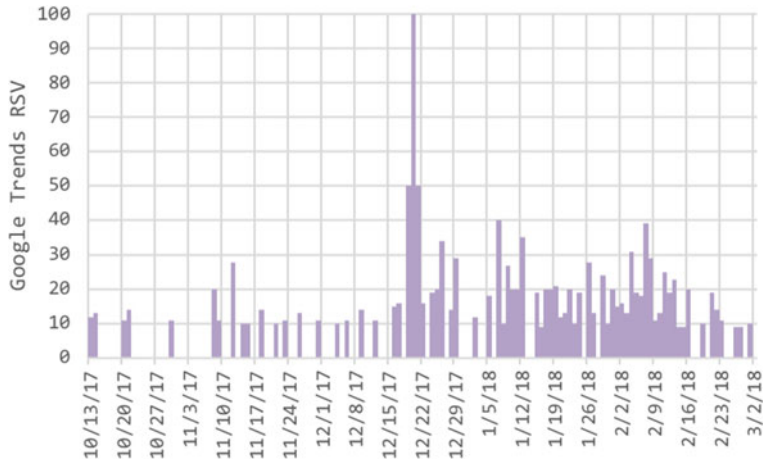

Fig. 28.2 Daily Google Trends relative search volume (RSV) fractions from October 13, 2017—March 2, 2018. Days with zero search interest are blank

Table 28.2 Google Trends (GT) relative search volume pre– and post–media day event on December 20, 2017

	Number of days in time frame	% of days in time frame (%)	Sum of GT search fractions	% of fractions (%)	Number of non-zero search days	% non-zero search days (%)
Pre-event: 10/13/17–12/19/17	68	48	326	23	22	32
Event: 12/20/17	1	1	100	7	1	100
Post-event 12/21/2017–3/2/2018	72	51	1013	70	53	74

Figure 28.2 shows the daily relative search volume fractions for the term "hepatitis". A strong spike in volume was noted on December 20, 2017 and the search volume remained elevated for two months.

28.3.2 R_0, R_{obs}, *and Final Outbreak Size Estimates*

The basic reproduction number (R_0), observed reproduction rate (R_{obs}), and final outbreak size estimates are listed in Table 28.3. These estimates were calculated using the IDEA model with inputs from WHO, HM, and GT + HM sources.

Table 28.3 Basic reproduction number (R_0), observed reproduction number (R_{obs}), and final outbreak size as estimated using case counts from WHO, HM, and GT + HM data and the IDEA model. GT + HM = HealthMap data smoothed with Google Trends relative search volume data; HM = HealthMap data; WHO = World Health Organization data

Basic Reproduction Number: R_0			
Data Source	*Mean*	*Min*	*Max*
WHO	1.58	1.32	1.87
HM (raw)	1.57	1.32	1.85
GT + HM	1.97	1.55	2.47
Observed Reproduction Number: R_{obs}			
Data Source	*Mean*	*Min*	*Max*
WHO	1.27	1.17	1.37
HM (raw)	1.28	1.18	1.40
GT + HM	1.19	1.11	1.28
Final Outbreak Size			
Data Source	*Mean*	*Min*	*Max*
WHO	1858	1458	2402
HM (raw)	2463	2008	2897
GT + HM	930	918	944

28.4 Discussion

There is growing evidence in the literature that non-traditional digital surveillance data can accurately estimate transmission dynamics, including case counts and basic and observed reproduction numbers, during an acute infectious outbreak (Ocampo et al. 2013; Yang et al. 2017; Alicino et al. 2015; Yang et al. 2017; Majumder et al. 2016) Our analysis supports the hypothesis that non-traditional data sources such as cumulative GT RSV data and HM data correlate well with traditional epidemiological surveillance data sources such as WHO cumulative case counts.

The basic and observed reproduction number estimates estimated by the IDEA model for the WHO, HM, and GT + HM data are similar across the three data sources and consistent with previously published R_0 values from a Hep-E outbreak in Uganda in 2007–2009 (Nannyonga et al. 2012, Nishiura 2019). This was true even in Namibia which has a low prevalence of hepatitis-E, a low internet access rate, and in the face of a media event which affected search volume.

Final outbreak size estimated using the GT + HM data was smaller compared to HM and WHO. The case count curves between the 3 data sources correlated well; however, the GT + HM data demonstrated a slowing of search volume in March 2018 despite continued progression of the outbreak. This resulted in the GT + HM estimate for final outbreak size to be smaller than the HM and WHO estimates, both of which were closer to outbreak sizes reported in August 2018 (Nkala 2018). Hepatitis E is a disease that has relatively low mortality and most patients are asymptomatic or they recover fully (with a small number of tragic exceptions, especially among pregnant women) (Center for Disease Control and Prevention 2018). Unlike Ebola or Zika, which drive large search volume out of fear, worry, or fascination, Hep-E

does not dominate news cycles. In this context, the HM data, as a non-traditional source of outbreak data, may be a more useful tool in predicting final outbreak size than GT-derived data sources (e.g. GT + HM).

Of note, our data was collected only during the initial phase of the outbreak, though transmission persisted through 2018 (Nkala 2018). Notably, at time of analysis, only the first five months of "ground truth" data were available from the WHO, likely due to limited public health resources, strained medical infrastructure, and limited laboratory capability. In this setting, the combination of GT and HM may be a useful adjunctive source of information to model transmission dynamics and guide public health responses.

However, to compare across data sources, the HM and GT + HM data sources were artificially truncated in this paper. More accurate estimates of transmission dynamics may be possible when applied to a full data set for the entire outbreak; this said, public health officials and government officials must often make decisions early in the outbreak without the benefit of a complete, accurate data set, and as such, the analytical approach highlighted here may be useful even under such circumstances. Public health and government officials who are tasked with responding to an acute infectious outbreak need near real-time, accurate information about the status and characteristics of an outbreak, especially transmissibility, to plan effective intervention strategies and to deploy resources effectively. Whether used as a supplement to traditional epidemiological data sources in middle and high resource settings, or as a stand-alone data source in low resource settings, nontraditional data sources may be a useful tool to aid in the fight against acute infectious outbreaks.

Conflicts of Interest: The authors have no conflicts of interest.

Author Contributions Michael Morley, Maiamuna S. Majumder, Tony Gallanis, and Joseph Wilson participated in conceptualization, data curation, formal analysis, validation, writing the original draft, reviewing and editing. Maiamuna S. Majumder also obtained the HM data and participated in the supervision of the research team and the project.

References

Alicino, C., Bragazzi, N., Faccio, V., Amicizia, D., Panatto, D., et al. (2015). Assessing Ebola related web search behavior: insights and implications from an analytical study of Google trends-based query volumes. *Infectious Diseases of Poverty 4*, 54 https://doi.org/10.1186/s40249-015-0090-9.

Center for Disease Control and Prevention. (2018). Center for Disease Control and Prevention Hepatitis E-FAQs. (Revised, May 9, 2018). Retrieved October 12, 2018, from https://www.cdc.gov/hepatitis/hev/hevfaq.htm.

Fisman, D. N., Hauck, T., Tuite, A., Greer, A. L. (2013). An idea for short term outbreak projection: Nearcasting using the basic reproduction number. *PLoS One*, *8*(12), e83622. https://doi.org/10.1371/journal.pone.0083622.

Google Trends. (2018). Retrieved April 25, 2018, from https://support.google.comtrends/answer/4365533?hl=en.

HealthMap. (2019). Retrieved June 23, 2019, from https://www.healthmap.org/en/.

Majumder, M. S., Santillana, M., Mekaru, S.R., McGinnis, D.P., Khan, K., & Brownstein, J.S. (2016) Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015–2016 Colombian Zika virus disease outbreak. *JMIR Public Health Surveill, 2*(1), e30 https://publiche.jmir.org/2016/1/e30.

Nannyonga, B., Sumpter, D. J. T., Mugisha, J. Y. T., & Luboobi, L. S. (2012). The dynamics, causes and possible prevention of Hepatitis *E* Outbreaks. In Y.E. Khudyakov, ed. *PLoS ONE*, 7(7), e41135. https://doi.org/10.1371/journal.pone.0041135.

Nishiura, H. (2019). Household data from the ugandan Hepatitis E Virus outbreak indicate the dominance of community infection. *Clinical Infectious Diseases*, *51*(1), 117–118. (1 July 2010). https://doi.org/10.1086/653448 Retrieved May 23, 2019, from https://academic.oup.com/cid/article/51/1/117/297883.

Nkala, O. (2018). Hepatitis-E death toll rises to 24 in Namibia, outbreak news today, (August 27, 2018) http://outbreaknewstoday.com/hepatitis-e-death-toll-rises-24-namibia-52443/.

Nuti, S., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R., et al. (2014). The use of Google trends in health care research: A systematic review. *PLOS One, 9*(10), e109583. Retrieved April 25, 2018 from https://doi.org/10.1371/journal.pone.0109583.

Ocampo, A. J., Chunara, R., & Brownstein, J. S. (2013). Using search queries for malaria surveillance Thailand. *Malaria Journal, 12,* 390. https://doi.org/10.1186/1475-2875-12-390.

World Health Organization. (2018a). World Health Organization Outbreak News Hepatitis-E Namibia. (January 15, 2018) https://www.who.int/csr/don/15-january-2018-hepatitis-e-namibia/en/.

World Health Organization. (2018b). World Health Organization Fact Sheet Hepatitis-E. (September 19, 2018). Retrieved October 12, 2018, from http://www.who.int/news-room/fact-sheets/detail/hepatitis-e.

World Health Organization. (2018c). Africa Weekly Bulletin on outbreaks and other emergencies Retrieved April 25, 2018, from https://www.afro.who.int/publications/outbreaks-and-emergencies-bulletin-week-51-16-22-december.

Yang, S., Santillana, M., & Kou, S. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *PNAS*, *112*(47), 14473–14478. Retrieved April 2018 https://doi.org/10.1073/pnas.1515373112.

Yang, S., Kou, S. C., Lu, F., Brownstein, J. S., Brooke, N., & Santillana, M. (2017). Advances in using Internet searches to track dengue. *PLoS Computational Biology, 13*(7), e1005607. https://doi.org/10.1371/journal.pcbi.10056.